AD_____

Award Number: W81XWH-10-1-0790

TITLE: Global Genomic Analysis of Prostate, Breast and Pancreatic Cancer

PRINCIPAL INVESTIGATOR: Dr. Richard M. Myers

CONTRACTING ORGANIZATION: HudsonAlpha Institute for Biotechnology
Huntsville, AL 35806-2908

REPORT DATE: October 2012

TYPE OF REPORT: Addendum to Final

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE	**2. REPORT TYPE**	**3. DATES COVERED**
29 October 2012	Addendum to Final	16 September 2010- 15 September 2012

4. TITLE AND SUBTITLE
Global Genomic Analysis of Prostate, Breast and Pancreatic Cancer

5a. CONTRACT NUMBER

5b. GRANT NUMBER
W81XWH-10-1-0790

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
Katherine E. Varley, Marie K. Cross, Richard M. Myers

E-Mail: rmyers@hudsonalpha.org

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
HudsonAlpha Institute For Biotechnology
Huntsville, AL 35806-2908

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for Public Release; Distribution Unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

Prostate cancer and breast cancer are the most prevalent cancers in men and women, respectively, and pancreatic cancer, while more rare than prostate and breast cancer, is an extremely lethal cancer, with a 5-year survival rate of less than 5%. With the recent advances in next generation sequencing technology, there is an opportunity to identify genomic aberrations that will provide knowledge about the biology driving these cancers and serve as novel diagnostic and prognostic biomarkers for these diseases. We performed genome-wide measurements of mRNA, microRNA, and DNA methylation from tumor and patient-matched non-tumor tissues. We have identified DNA methylation and gene expression signatures associated tumor formation, disease recurrence, and treatment sensitivity. These candidate biomarkers may be useful in predicting clinical outcomes for patients, and may suggest pathways that can be targeted with novel treatment regimens.

15. SUBJECT TERMS- None provided

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON**
a. REPORT	**b. ABSTRACT**	**c. THIS PAGE**			USAMRMC
U	U	U	UU	50	**19b. TELEPHONE NUMBER** *(include area code)*

Table of Contents

Final Report for W81XWH-10-1-0790

TITLE: Global Genomic Analysis of Prostate, Breast and Pancreatic Cancer

PRINCIPAL INVESTIGATOR: Dr. Richard M. Myers

Introduction

The goal of this project is to use state-of-the-art functional genomics assays to provide an unprecedented depth of information about the molecular defects that occur in three types of cancer that occur frequently in military personnel, specifically prostate, breast and pancreatic cancer. With a more comprehensive and detailed understanding of these cancers, we hope to decrease the burden of these diseases in military personnel and the general population by identifying biomarkers that will increase the rate and sensitivity of early detection and that predict which treatments will be most effective in certain subtypes of disease, and by identifying pathways and molecular defects that can be targeted by novel therapeutics. To achieve these goals, we are working with collaborators at Stanford University and the University of Alabama at Birmingham, who are providing de-identified frozen tumor specimens, along with detailed clinical data, from individuals with these cancers. After macrodissection of each sample to provide separated tumor and nearby non-tumor tissues, along with pathological analyses to ensure that each type is indeed separated, we extract nucleic acids (RNA and DNA) from each sample, while saving a portion of the intact tissues if enough is present. We then perform multiple functional genomic assays, as well as detailed genetic analyses, by new "next-gen" DNA sequencing methods in an effort to gain key insights into the molecular defects that contribute to these cancers. These assays include DNA methylation analysis, mRNA profiling, and microRNA profiling. When analyzed together, these data can provide a broad picture of the gene regulatory defects present in tumors that can be used to classify disease states and identify pathways that are frequently altered during carcinogenesis. We are making significant headway in implementing this large-scale analysis of clinical samples, and our earliest analyses reveal striking tumor-specific molecular defects. Analysis of these large genome-wide datasets is ongoing and we are confident that we will continue to identify novel clinically relevant molecular features in these diseases.

Body

Our research program, supported by W81XWH-10-1-0790, involves applying new high-throughput genomic techniques to increase our understanding of the molecular basis of three types of cancer that are of critical importance to the military, their families, and civilians. We are applying these approaches to study the genetics and genomics of prostate, breast and pancreatic cancer, as well as to the differential responses of women with breast cancer to new experimental drug treatments. Our goals are to identify genetic, epigenetic and genomic biomarkers for these diseases that can be used for accurate and subtype-specific diagnosis, for prognosis of individuals at any stage of the diseases, for determining the most effective treatment regime, and, ultimately, for the development of new and more effective therapies. To accomplish these goals, we are developing and applying new ultrahigh-throughput DNA sequencing technologies ("next-gen sequencing") to probe at an unprecedented level of detail and comprehensivity the genomes of the tumors and matching non-tumor tissues from individuals with these cancers. During this project, we are using these approaches to measure,

on a genome-wide scale, the mRNA, microRNA, DNA methylation in tumors and non-tumor tissues from at least 50 individuals with each cancer.

A brief summary of this progress is listed here. In Year 1, we: 1) obtained IRB approval and material transfer agreements, and identified clinical sources, for the tissues needed for the project from our two major collaborating institutions; 2) developed protocols and the infrastructure to macrodissect the tumor and non-tumor tissues from study participants; 3) began macrodissection on the breast and pancreatic cancer samples; 4) performed DNA methylation profiling in prostate cancer and breast cancer; 5) developed protocols and performed mRNA profiling on 28 breast cancer cell lines; 6) developed and quality-control tested an optimized method for measuring and interpreting microRNAs; 7) began development of a protocol for performing our genomic assays in formalin-fixed paraffin embedded (FFPE) tissues, which will allow us to access larger collections of relevant patient samples and bypass obstacles presented by the macrodissection of frozen tissues; and 8) designed and tested a suite of computational algorithms that allow efficient analysis of DNA methylation profiles as measured by sequencing (the RRBS method described in our proposal).

In Year 2, we: 1) homogenized and isolated nucleic acids from 79 pancreas and 86 breast samples; 2) constructed RRBS libraries for pancreas and breast cancer; 3) constructed RNA-seq libraries for prostate, pancreas and breast cancer; 4) identify DNA methylation signatures associated with subtype in breast cancer, biochemical recurrence in prostate cancer and novel tumor subtypes in pancreatic cancer; 5) developed computational pipelines to align the 50 million RNA-seq reads from each sample to the genome, assign them to transcripts or compile reads at un-annotated regions into a novel gene prediction, and produce normalized expression values for each gene; 6) analyzed gene expression signatures associated with subtype in breast cancer, biochemical recurrence in prostate cancer and tumorigenesis in pancreatic cancer; 7) performed microRNA-seq on 28 breast cancer cell lines; 8) implemented a computer program to identify fusion genes from RNA-seq data and identified read-through fusion transcripts significantly associated with breast cancer, as well as a fusion transcript significantly expressed in pancreatic tumors.

1) Year 1

a) Sample accrual and macrodissection

In Year 1 we identified, obtained and macrodissected clinical specimens for each cancer type:

i) Prostate cancer: Our collaborator at Stanford University, Dr. James Brooks, a urologist, surgeon and expert in prostate and kidney cancer genomics, has provided us with genomic DNA from 73 macrodissected prostate tumor tissues and 63 normal prostate tissues, mostly from the same individual patients. We also obtained total RNA for our mRNA and microRNA analyses from a subset of these individuals (70 tumor samples and 36 normal samples).

ii) Breast cancer: For the breast cancer retrospective study of young women (<50 years old) with ER+ HER2- breast cancer, we have received 72 samples from 48 patients.

iii) Breast cancer clinical trial: We obtained tumor biopsy specimens from a breast cancer clinical trial. This trial is a study of the clinical response of ER+ breast cancer to treatment with a combination of letrozole and bevacizumab. We received 32 tumor biopsies obtained at different time-points during the course of treatment of 19 patients that can be used to monitor the molecular response to treatment. This study will be the first of its kind to perform next-gen functional genomics assays from biopsies obtained during a clinical trial, an extremely important

goal for understanding and predicting which patients will respond to this treatment regime in the future.

iv) Pancreatic cancer samples: Pancreatic cancer is often diagnosed in late stages of the disease, and tumor resection is a high-risk procedure, so collections of research tissue specimens are exceedingly rare. We have identified a cohort of 100 frozen pancreatic tissue specimens through our collaborators at UAB. After macro-dissection we received 52 tumors and 27 matched non-tumor tissues.

v) We obtained 28 breast cancer cell lines from our collaborators at UAB that were established from patient tumors and represent diverse breast cancer subtypes. These samples can be used to define molecular profiles of breast cancer subtypes without the residual adjacent normal cell populations that remain even after careful macrodissection of primary tumors. Furthermore, these cell lines have been evaluated for sensitivity to several chemotherapeutic regimes, and the molecular features that correlate with response to treatment in this pre-clinical model will likely be relevant in primary tumors as well.

b) Progress on DNA methylation profiling

We are using a combination of two complementary methods to obtain detailed DNA methylation profiles of the samples in all of our studies. These are Reduced Representation Bisulfite Sequencing (RRBS; Meissner et al., 2008), which uses next-gen sequencing of bisulfite-treated genomic DNA, and a method from Illumina called the Methyl450 assay, which also relies on bisulfite treatment, but uses single-base sequencing on an array to quantitate methylation at specific CpG residues. This combination of RRBS and Methyl450 is valuable because the methods measure, at a quantitative level, the fraction of methylation at 700,000 and 450,000 CpGs, respectively, with only a very small amount of overlap (fewer than 4%) between the two methods. Further, RRBS can detect genetic variants as well as non-CpG cytosine methylation in the regions it assays, and the Methyl450 assay tests more than 3,000 genes not covered by RRBS (Sandoval et al., 2011). Together, these assays provide a broad picture of the methylation differences between tumor and normal tissues, and allow us to detect specific CpGs with disease-associated methylation changes.

i) Breast cancer cell line methylation

We spent part of Year 1 improving the DNA methylation profiling methods, simplifying and hardening the library preparation steps, and learning to multiplex samples in the next-gen sequencing step to decrease costs and increase throughput. We performed RRBS on 28 breast cancer cell lines, and obtained quantitative DNA methylation for 796,861 CpG loci across the genome. We found that a large subset of CpGs have highly variable DNA methylation across these breast cancer cell lines. We performed unsupervised hierarchical clustering of 10,000 CpGs with the most variable DNA methylation across cell lines, and found that these CpGs divided the cell lines into clades/clusters that correlated with Luminal or Basal subtypes as defined by gene expression and immunohistochemistry (see Figure 1 in Supporting Data, at the end of this document). We continued to use this dataset in the breast cancer analyses performed in Year 2 which involved integrating them with other molecular, phenotypic, and pharmacological response data to better understand the consequences of these drastically different methylation profiles between breast cancer cell lines.

ii) Prostate cancer DNA methylation

During Year 1, we used the Methyl450 assay to quantitate DNA methylation in our initial set of prostate cancer samples (73 primary prostate tumor tissues and 63 benign-adjacent prostate

tissues). Our analysis so far has shown that tumor and normal tissues can be easily separated based on DNA methylation patterns. Principal Component Analysis (PCA) performed on the approximately 450,000 CpGs distinguishes most normal tissues from tumor tissues, and suggests that the DNA methylation patterns between tumor tissues are more diverse than DNA those patterns between normal tissues, as seen from the increase in distances between each tumor sample in the PCA plot compared to the normal samples (Figure 2, Supporting Data). Unsupervised hierarchical clustering of the data supports the PCA data; we observed two main clusters, one consisting mostly of tumor tissues (64T/1N), and one consisting mostly of normal tissues (62N/9T). We discovered ~220,000 CpGs that have significantly different methylation patterns between normal and tumor prostate tissue samples by using several different parametric and nonparametric statistical analyses. The majority of these aberrantly methylated CpGs (66.4%) are hypermethylated in the tumor tissue samples compared with the normal tissue sample. Focusing on the CpGs with 10% or greater standard deviation, ~37,000 CpGs had methylation scores that were significantly different between normal and tumor. Chi-squared analysis suggests that there is no specific region of the genome where these significant CpGs are located, but rather, CpGs that are differentially methylated between normal and tumor prostate tissue are distributed throughout the genome. Future analysis will determine other characteristics of these significant CpGs, such as whether they are preferentially located within CpG islands.

An important question in prostate cancer biology is whether specific genomic or genetic features drive a prostate tumor to be aggressive. Unsupervised hierarchical clustering of the 12,000 most variable CpGs within the tumor tissues alone show distinctive clusters with pockets of unique methylation patterns (Figure 3, Supporting Data). We are interested in understanding how these different methylation patterns within the tumor population correlate with the clinical data associated with these samples, such as the patients Prostate Specific Antigen (PSA) score and the Gleason grade of the tumor. Preliminary statistical analysis on the prostate tumor samples with integrations of downstream clinical data suggests that the methylation patterns of CpGs within the coding regions of a subset of genes (i.e., in the gene "bodies", as opposed to the flanking regions of genes) correlate with specific clinical covariates. We are currently continuing to integrate clinical data into our statistical analyses to understand what molecular differences may be associated with distinct prostate tumor phenotypes.

c) Progress in mRNA profiling

As we discussed in our initial application, the use of next-gen sequencing to analyze mRNA gene expression patterns (RNA-seq) has many substantial advantages over microarray methods that use hybridization. RNA-seq is much more accurate, providing quantitative measures of mRNAs over a very wide dynamic range. It identifies both known and unknown mRNAs, as well as all isoforms, including different spliced versions, different 5' ends, and different 3' ends. It identifies allele-specific or allele-biased mRNA expression, as well as versions of mRNAs that have been edited. We have used RNA-seq in a variety of projects, and we and others have found many examples of new findings that were not observed with older methods. Thus, RNA-seq is an important method to apply to the cancer studies we are performing in this project.

The mRNA-seq protocol typically used in our laboratory requires 5 μg total RNA, using a protocol developed as part of the ENCODE Project by our collaborators led by Dr. Barbara Wold (Mortazavi et al., 2008). Early in this TATRC cancer project, we realized that the total RNA yields from our macrodissected tumor specimens were likely to be too low for such a large requirement. Thus, during Year 1 of this project as well as part of our ENCODE work, we worked on protocols for mRNA-seq that could provide accurate and complete quantification of

the transcriptome from smaller amounts of starting total RNA. We developed a new method for making next-gen sequencing libraries for mRNA-seq from very tiny amounts small amounts of total RNA (Gertz et al., 2011, Appendix A). We tested this new protocol on the 28 breast cancer cell lines. During this test, we learned that to obtain reproducible measurements of gene expression from small quantities of total RNA (as low as 10 ng), it was critical to normalize the amount of cDNA used in the library construction. We also used this dataset to begin developing the bioinformatics pipeline for analyzing mRNA-seq from cancer samples, including the implementation of an analytic approach for detecting viral RNA in human cancer samples. The results of the preliminary data analysis are described below, demonstrating that mRNA-seq can accurately distinguish breast cancer subtypes.

From producing and analyzing mRNA-seq data on the 28 breast cancer cell lines, we obtained gene expression measurements for 32,062 human transcripts. In all our RNA-seq experiments, we collect at least 25 million pairs of 50bp next-gen sequencing reads per sample to ensure reproducible measurements with a large dynamic range to quantitate subtle expression differences. We determined the correlation coefficient between the expression values for each pair of samples, and then clustered the samples based upon their correlations to each other (Figure 4, Supporting Data). We found that the samples clustered into major clades, and were representative of different breast cancer subtypes, with luminal samples clustering together, and basal samples clustering together. Interestingly, these gene expression clusters indicate that there are subgroups within the basal and luminal subtypes, and we are investigating these further to determine whether these divisions correlate with other molecular, phenotypic or pharmacologic parameters. We also developed quality control metrics for each library, including checking for concordance between RNA-seq expression values for the HER2, ER, and PR genes and immunohistochemical measurements of these standard biomarkers. In addition to aligning to human transcripts, we also developed a bioinformatics pipeline to detect viral mRNA that may be expressed in tumors. We were able to detect two viruses, SV40 and a polyoma virus, in two different cell lines. We are further investigating these viruses and their possible origins, and believe that this tool will be valuable for determining whether primary tumors from patients contain these viruses. This collection of mRNA-seq data from breast cancer cell lines has been a great test-set as we expand our informatics pipeline to include the detection of novel transcripts, SNPs and mutations.

Once we were satisfied that our protocol was robust and could be use with the small quantities of total RNA obtained from macrodissected tumor specimens, and found that it accurately distinguishes cancer specific differences in gene expression, we began performing mRNA-seq for our first prostate cancer cohort (73 prostate tumors and 63 non-tumor controls). Given the exciting findings that we have obtained with RNA-seq from a modest number of cell lines, we believe that it is likely that examining these larger sets of tumors and non-tumor samples will provide new insights into each type of cancer.

d) Progress in microRNA profiling

It is clear that microRNAs, while a relatively recent discovery, are key regulators and play important roles in a large number of biological processes. As part of this TATRC cancer project, we are measuring microRNAs in each of the tumor and non-tumor samples for each of the cancers we are studying.

During the first part of Year 1, we tested a variety of protocols for generating microRNA libraries for next-gen sequencing, and eventually developed our own protocol. We focused first on making libraries that would allow one microRNA sample to be sequenced per "lane" on the Illumina next-gen sequencing machine. Sequencing several such libraries showed that this

method is accurate and reproducible at quantifying microRNA levels from human tissues. There are ~1,733 identified human microRNA in the mirBase database, and, with our library and sequencing method, we are detecting ~700-900 unique known microRNAs per library, depending on tissue type (Langmead et al., 2009). In early in Year 2, we improved this method so that we can barcode and multiplex-sequence several samples per Illumina sequencing lane, thus reducing costs for microRNA sequencing.

e) Development of methods to use FFPE tissues for our cancer genomics studies

The vast majority of cancer specimens obtained in hospitals across the country are preserved as Formalin-Fixed Paraffin Embedded (FFPE) specimens for pathological examination, and are more readily available for research than frozen samples. An additional benefit is that FFPE specimens can be dissected at room temperature without the expensive cryo-preservation equipment or specialized pathology training required for frozen dissection. Unfortunately, this method of perseveration, while well-suited to classic pathology and immunohistochemistry, causes fragmentation of the DNA and RNA used in genomics studies (Medeiros et al., 2007). In Year 1 we investigated whether the fragmentation of the DNA and RNA is compatible with our current assay protocols. We obtained five FFPE tumor specimens, including breast, prostate and pancreatic cancer, and tested DNA and RNA extraction from these tissues. We also tested RRBS, mRNA-seq and microRNA-seq library construction on these specimens. We obtained very preliminary data that suggested our protocols were compatible with fragmented DNA and RNA from FFPE specimens should we pursue acquisition of these types of samples.

f) Development and improvement of computational algorithms for analyzing genomic and genetic data based on next-gen sequencing

Before this TATRC project started, we had built an automated computational pipeline to collect, transfer, and store next-gen DNA sequencing data and perform the initial primary analysis on them, including base calling, quality score determination, alignment to the genome sequence, and basic quality control metrics. While we have continually upgraded this part of the pipeline, particularly as our throughput has increased, it is stable and robust, and can handle the large datasets that we are generating for this and other projects. However, the downstream, biological interpretation of the data types that we are generating in this project require much more extensive and complex analysis, and each data type requires specifically-designed analysis tools. These algorithmic suites not only needed to be developed, but have required hardening and automation to handle the very large datasets that we are generating for this TATRC project. We spent efforts in Year 1 developing many of these methods, including algorithms for calling and quantitating methylation data from RRBS experiments and methods for analyzing mRNA-seq data for these projects. The methylation methods are now robust and part of our regular pipeline, and we have the beginnings of an automated method for measuring mRNAs and microRNAs. However, these latter data types will require some more development of the analysis pipeline, particularly to allow the quantitation of different mRNA isoforms by RNA-seq. This development is ongoing and will likely require another half-year before they are fully implemented.

2) Year 2

In Year 2, we learned that we would not receive funding for subsequent years, so we focused our efforts on processing and analyzing the samples and data for aspects of the project that were initiated in Year 1.

a) Genomic Analysis of Breast Cancer

In Year 1 we sequenced the transcriptome (RNA-seq) to measure gene expression differences between the 28 breast cancer cell lines and performed reduced representation bisulfite sequencing (RRBS) to detect genome-wide DNA methylation differences between the 28 breast cancer cell lines. In Year 2 we developed computational pipelines to align the 50 million RNA-seq reads from each sample to the genome, assign them to transcripts or compile reads at un-annotated regions into a novel gene prediction, and produce normalized expression values for each gene. We also refined our computational analysis of DNA methylation to identify regional differences in methylation, in addition to querying individual CpGs for significance. The panel of 28 breast cancer cell lines that we first used this analysis on includes several subtypes including TNBC basal A, TNBC basal B, HER2 positive basal-like, HER2 luminal-like and ER+ luminal breast cancer cell lines. The genome wide functional genomics data in these cell lines allows us to identify differences between subtypes of breast cancer and to identify genomic signatures associated with sensitivity and resistance to various therapeutics that have been tested on these cell lines. We used linear regression to identify DNA methylation and gene expression differences that were associated with sensitivity to 75 different therapeutic compounds, quantified as IC50 values, from several recent publications (Oliver et al., 2012, Heiser et al., 2012). Strikingly, the same gene sets were associated with response to multiple drugs, indicating common pathways of resistance to different types of chemotherapy. We are further investigating these pathways to determine if combinations of different therapeutics that target resistance associated pathways can lead to increased sensitivity. In particular we were interested in determining genomic signatures that are associated with sensitivity to TRA-8, a monoclonal antibody to the death receptor developed by our collaborators at UAB that is effective at reducing cell proliferation in basal breast cancer. We identified 456 CpGs whose DNA methylation was significantly associated with sensitivity to TRA-8 (FDR < 0.05), and 328 genes whose expression was significantly associated with sensitivity to TRA-8 (FDR < 0.05). Both of DNA methylation and gene expression associated with TRA-8 sensitivity occur at genes that are involved in cell adhesion ($p < 3.39e-02$). We are investigating the role of cell adhesion molecules in modifying the accessibility of cell surface receptors to antibody therapy.

The genome-wide analysis of gene expression and DNA methylation allowed us to investigate other molecular differences across the 28 diverse cell lines. Paired-end RNA-seq data provides the opportunity to identify fusion genes, which play a significant role in the development and progression of several types of cancer, particularly hematological malignancies. To determine if the breast cancer cell lines harbor fusion genes, we analyzed the RNA-seq data with ChimeraScan, a computer program designed to detect fusion genes from RNA-seq data by identifying transcripts containing sequence from two different genes (Iyer et al., 2011). From this analysis we determined that RNA transcripts composed of sequences from two adjacent genes occur frequently in breast cancer cell lines. Read-through fusion transcripts such as these were recently associated with prostate cancer progression, so we focused our analysis on this type of defect in breast cancer. We determined how many of these fusion transcripts were detected in our other RNA-seq datasets from TNBC primary tumors, and estrogen receptor positive primary tumors, as well as normal control tissues. We identified three read-through fusion transcripts that were significantly associated with breast cancer and that occur frequently across breast cancer samples. Western blots performed on the breast cancer cell lines harboring the fusion transcripts suggest that they are translated into fusion proteins. We have prepared a manuscript describing this discovery (Appendix B).

The genome-wide DNA methylation and gene expression data from diverse breast cancer cell lines also provides the opportunity to explore gene regulatory mechanisms responsible for the large-scale gene expression differences that define breast cancer subtypes. Targeting the master regulators responsible for the transcriptional programs in each subtype could be an

effective targeted therapy. We found that genome-wide DNA methylation signatures recapitulate the breast cancer subtype classifications. We hypothesized that intergenic loci that are specifically unmethylated in each subtype are active regulatory regions, and that transcription factors binding to these loci are master regulators of the subtype-specific gene expression signatures. To test this hypothesis, we compared intergenic loci that are specifically unmethylated in luminal breast cancer cell lines with hundreds of ChIP-seq datasets publicly available from the ENCODE Project. Regulatory regions that are specifically unmethylated in luminal breast cancer cell lines were significantly enriched for transcription factor binding sites for estrogen receptor and its cofactors FOXA1 and GATA3. These transcription factors are known to be master regulators of the luminal gene expression signature and are highly effective drug targets for luminal breast cancer. We repeated this process to identify potential master regulators in TNBC cell lines, a subtype with no effective targeted therapy. We found that glucocorticoid receptor (GR) and STAT3 transcription factor binding sites were enriched at unmethylated regulatory regions associated with the TNBC expression signature. We are performing ChIP-seq in TNBC cell lines and luminal cell lines to confirm the binding the transcription factors GR and STAT3 near genes with TNBC-associated expression. Four cell lines of each subtype were used to prepare dexamethasone-induced and untreated control cells for these studies. We are also testing small molecule inhibitors of both factors to determine whether they specifically inhibit TNBC cell proliferation. Representative cell lines were selected from TNBC (basal A and B) and luminal subtypes for in vitro cell viability assays to assess the effects of inhibitors of STAT3 (SI3-201) or GR (mifepristone). These transcription factors, GR and STAT3 could be promising therapeutic targets for inhibiting the transcriptional network driving TNBC proliferation. This investigation demonstrates the power of integrating data from genome-wide functional genomics assays to uncover the master regulators of gene expression signatures associated with breast cancer subtypes.

In Year 1 we had obtained frozen tumor tissue specimens from 48 young women with ER+ HER2- breast cancer. We also obtained frozen tumor tissue specimens from 19 post-menopausal women with ER+ HER2- breast cancer. In Year 2 we successfully isolated nucleic acids from these samples and performed RNA-seq and RRBS on the samples. We are in the process of comparing these samples to identify gene expression and DNA methylation signatures that differ between ER+ HER2- breast tumors that occur in pre and post-menopausal women. We are also comparing these samples to 46 triple negative breast cancer (TNBC) tumors that we analyzed as part of another study. This comparison will allow us to identify age related molecular differences in breast cancer that are independent of subtype. In addition we have analyzed the DNA methylation differences between ER+ primary tumors and TNBC primary tumors to determine if the subtype specific methylation signatures that we identified in the breast cancer cell lines are recapitulated in primary tumors. We confirmed that the subtype associated methylation differences in primary tumor support the hypothesis that STAT3 and GR binding sites are differentially methylated between subtypes, and demonstrated that ER+ primary tumors have overall significantly more methylation. We are further investigating the global hypermethylation of ER+ primary tumors to determine how similar the methylated loci are to those found in the hypermethylator phenotype reported in colorectal cancer and gliomas. This is an exciting finding and suggests there are global epigenetic differences between these types of breast cancers, and may indicate that DNA methylation inhibitors may be affected therapeutic strategies in this subset of breast cancers.

b) Genomic Analysis of Prostate Cancer

In Year 2 of the TATRC project, we continued our analysis of the DNA methylation patterns in prostate cancer. One important question in the prostate cancer field is whether there are

biomarkers that can distinguish aggressive prostate cancer from non-aggressive prostate cancer. We integrated the DNA methylation data from the prostate tumor tissues with patient clinical information using linear regression models, in order to determine whether there were methylation signatures that were associated with specific patient clinical information, such as Gleason grade of the tumor or patient PSA score. We discovered methylation patterns at specific CpG loci that associated with biochemical recurrence in a linear regression model. Biochemical recurrence is defined as a rise in the level of prostate specific antigen (PSA) in the bloodstream after a radical prostatectomy, and is a first indicator of potential clinical recurrence of prostate cancer. Currently, clinicians use Gleason grade to assess prognosis of the disease and risk assessment tools such as CAPRA-S (Cancer of the Prostate Risk Assessment Post-Surgical), developed at UCSF, to determine whether a patient has a high likelihood of undergoing recurrence after radical prostatectomy (Cooperberg et al., 2011). We tested these CpG methylation patterns in logistic regression and plotted the results in Receiver Operating Characteristic (ROC) curves, which are used to evaluate the sensitivity and specificity of diagnostic and prognostic tests by comparing the number of true positives versus the number of false positives that are classified at different thresholds (Zou et al., 1997). This analysis provided very promising results demonstrating that combining this methylation signature with patient clinical information sensitively and specifically anticipates which patients will biochemically recur. We are currently working towards validating these results in other prostate tissues, and we are continuing analysis of the prostate methylation patterns.

In Year 2 of the TATRC project, we initiated experiments to study RNA transcript expression patterns in prostate cancer. We constructed and sequenced RNA-seq libraries using RNA isolated from 81 of the136 prostate tissues used in the DNA methylation studies – 55 from prostate tumor tissues and 26 from benign-adjacent prostate tissues, representing 17 matched pairs. Resultant sequencing reads were run through the computational pipeline that we developed in Year 1, and we are currently analyzing this data. Thus far, out of the ~51,000 known coding and non-coding RNA transcripts in the genome, we have found ~500 transcripts that are significantly associated with biochemical recurrence. We are currently analyzing these transcripts to understand what cellular pathways are being enriched in these significantly expressed transcripts. We have also identified ~7,000 novel RNA transcripts across these prostate tissues. Novel transcripts are defined as transcripts that are expressed from un-annotated regions of the genome that have not previously been described as being regions containing genes. Through statistical analysis, we have found that ~70 of these novel transcripts are significantly associated with biochemical recurrence. We are particularly interested in the novel transcripts that are only being expressed in the prostate tumor tissues, as they may produce aberrant protein products that might serve as biomarkers for recurrence, or may serve as a novel target for prostate cancer therapeutics. We have also utilized a program called Chimerascan in order to identify chimeric transcripts in prostate cancer. When comparing the benign-adjacent tissues to the prostate tumor tissues, we confirmed the presence of the previously identified TMPRSS2-ERG fusion gene (Tomlins et al., 2005). We also identified other putative chimeras associated with prostate cancer using chimerascan, and we will be performing additional experiments to understand what role these fusions may play in prostate cancer etiology.

c) Genomic Analysis of Pancreatic Cancer

We originally proposed to study pancreatic cancer in Year 5 of the TATRC funding cycle, as given the progressive nature of this disease, identifying patient samples for this study is a challenge. However, in Year 1 of TATRC, we identified a cohort of 100 pancreatic tissues through our collaborations with the Pancreatic SPORE at the University of Alabama at

Birmingham. Given our advances in the efficiency and throughput of our genomic assays through technical development of our protocols, we decided to move the pancreatic cancer study forward in our TATRC timeline. In Year 1, one of our collaborators at the University of Alabama at Birmingham, Dr. William Grizzle, worked to macrodissect the pancreatic tumor tissue cohort in order to enrich for the tumor cell population. Early in Year 2 of TATRC, we obtained 52 macrodissected pancreatic tumor tissues, as well as 27 patient-matched uninvolved pancreatic tissues. These tissues were homogenized in our lab, and both DNA and RNA were extracted from the resultant lysate. After confirming the integrity of the nucleic acids through quality control assays, we constructed both RNA-seq libraries and Reduced Representation Bisulfite Sequencing libraries in order to study RNA expression and DNA methylation patterns in these pancreatic tissues.

Through analysis of the RRBS data, we have found that DNA methylation patterns can distinguish pancreatic tumor tissues from patient-matched uninvolved pancreatic tissues. Through statistical tests, we have discovered ~80,000 CpG loci that are significantly different between pancreatic tumor tissue and uninvolved pancreatic tissues, and we are currently working to understand the pathways that are being affected by these epigenetic alterations. Interestingly, when we cluster ~3,000 CpGs that have the most divergent methylation patterns between the tumor tissues and the uninvolved tissues, we see methylation patterns in the tumor tissues that may be indicative of pancreatic tumor subtypes.

Analysis of the RNA-seq data has uncovered that out of the ~51,000 known RNA transcripts in the human genome, ~20,000 of these transcripts have altered expression patterns in the pancreatic tumor tissues. Out of these ~20,000 RNA transcripts, ~55% are protein-coding RNAs, ~6% are small, regulatory non-coding RNAs, ~6% are long non-coding RNAs, and the remainder are pseudogenes and other non-coding transcripts. Analysis is ongoing to understand the cellular pathways that are enriched within these altered transcripts – we hope to uncover novel targets for pancreatic cancer therapeutics. Across the pancreatic tissues, we have discovered ~12,000 novel transcripts, being expressed from regions of the genome that are currently un-annotated. Of those ~12,000 novel transcripts, ~5,500 are significantly differently expressed between the pancreatic tumor tissues and the uninvolved pancreatic tissues. We are especially interested in the ~3,000 transcripts that are only expressed in the tumor tissues, as we hope these will ultimately express a protein that might be targetable by novel therapeutics for pancreatic cancer. We have also begun Chimerascan analysis of the pancreatic tissues, and we are excited to report that we have found two pancreatic tumor-specific chimeric transcripts. One of the pancreatic tumor specific chimeric transcripts that we have identified has previously been identified in both ovarian and breast cancer, so we are interested in performing additional experiments in order to understand how this particular transcript may be playing a role in pancreatic cancer.

Defining subtypes of different types of cancers, such as breast cancer, have proven important for informing clinical patient care. We see evidence for subtypes of pancreatic cancer in both the DNA methylation data and the RNA-seq data – specifically, we see evidence in our data that validate the three pancreatic cancer subtypes that were published by the Gray lab last year (Collisson et. al, 2011). We are continuing analysis to see if our RNA-seq data can aid in further clarifying and defining these subtypes of pancreatic cancer. Furthermore, we are working on integration of the DNA methylation data and the RNA-seq data, and we hope this will provide biomarkers for these different subtypes of pancreatic cancer, and potentially provide information on how these three subtypes might best be treated in the clinic.

Key Research Accomplishments

- identified sources and obtained IRB and institutional material transfer agreements for matched normal and tumor tissues for breast, prostate and pancreatic cancers.
- completed pathological analyses and macrodissection to enrich tumor cells from normal surrounding tissue for all three cancer tissue types.
- completed DNA methylation analysis on breast cancer cell lines, frozen prostate tissues, frozen breast tissues, and frozen pancreatic tissue.
- developed a high-throughput RNA-seq protocol.
- completed RNA-seq for breast cancer cell lines, frozen prostate tissues, frozen breast tissues, and frozen pancreatic tissue.
- developed a multiplex microRNA sequencing protocol, and performed microRNA-seq on 28 breast cancer cell lines.
- developed a suite of computational algorithms that allow efficient automated analysis of DNA methylation profiles, RNA-seq.
- identified DNA methylation and gene expression signatures associated with breast cancer cell line sensitivity to therapeutics
- identified 3 read-through fusion transcripts significantly associated with breast cancer in breast cancer cell lines and primary tumors
- identified gene expression and DNA methylation signatures associated with breast cancer subtypes that implicate transcription factors in controlling subtype specific transcriptional program that are potential therapeutic targets
- identified hypermethylation signature specific to ER+ breast cancer
- identified DNA methylation and gene expression signatures associated with biochemical recurrence in prostate cancer patients
- identified DNA methylation and gene expression signatures associated with distinct pancreatic cancer subtypes
- identified two pancreatic tumor specific chimeric transcripts

Reportable Outcomes

A manuscript that describes the improved library protocol for RNA-seq experiments was accepted for publication in Genome Research (Appendix A). We presented a poster describing the DNA methylation data and preliminary analysis from the prostate samples at the AGBT 2012 conference(Appendix C). We prepared and submitted a manuscript describing the discovery of read-through fusion transcripts in breast cancer (Appendix B).

Conclusions

The genomic analysis of breast, prostate and pancreatic cancer that we performed for this project led to several novel discoveries related to tumor formation, disease recurrence, and treatment sensitivity. The loci and pathways identified using these genome-wide approaches could serve as valuable diagnostic and prognostic biomarkers. In addition to predicting clinical outcome for individual patients, this information can lead to the identification of biochemical pathways that can be targeted by therapeutics that may lead to more effective treatments in the future. We are pursuing alternative funding mechanisms to support the further investigation of the promising biomarkers and candidate therapeutic targets identified in this study.

References

Christodoulou DC, Gorham JM, Herman DS, Seidman JG. Construction of normalized RNA-seq libraries for next-generation sequencing using the crab duplex-specific nuclease. Curr Protoc Mol Biol. 2011 Apr; Chapter 4: Unit 4.12.

Collisson EA, Sadanandam A, Olson P, Gibb WJ, Truitt M, Gu S, Cooc J, Weinkle J, Kim GE, Jakkula L, Feiler HS, Ko AH, Olshen AB, Danenberg KL, Tempero MA, Spellman PT, Hanahan D, Gray JW.Subtypes of pancreatic ductal adenocarcinoma and their differing responses to therapy. Nat Med. 2011 Apr;17(4):500-3.

Cooperberg MR, Hilton JF, Carroll PR. The CAPRA-S score: A straightforward tool for improved prediction of outcomes after radical prostatectomy.Cancer. 2011 Nov 15;117(22):5039-46.

Gertz J, Varley KE, Davis NS, Baas BJ, Goryshin IY, Vaidyanathan R, Kuersten S, Myers RM.Transposase mediated construction of RNA-seq libraries. Genome Res. 2012 Jan;22(1):134-41.

Gu H, Bock C, Mikkelsen TS, Jäger N, Smith ZD, Tomazou E, Gnirke A, Lander ES, Meissner A. Genome-scale DNA methylation mapping of clinical samples at single-nucleotide resolution. Nat Methods. 2010 Feb;7(2):133-6.

Heiser LM, Sadanandam A, Kuo WL, Benz SC, Goldstein TC, Ng S, Gibb WJ, Wang NJ, Ziyad S, Tong F, Bayani N, Hu Z, Billig JI, Dueregger A, Lewis S, Jakkula L, Korkola JE, Durinck S, Pepin F, Guan Y, Purdom E, Neuvial P, Bengtsson H, Wood KW, Smith PG, Vassilev LT, Hennessy BT, Greshock J, Bachman KE, Hardwicke MA, Park JW, Marton LJ, Wolf DM, Collisson EA, Neve RM, Mills GB, Speed TP, Feiler HS, Wooster RF, Haussler D, Stuart JM, Gray JW, Spellman PT.Subtype and pathway specific responses to anticancer compounds in breast cancer. Proc Natl Acad Sci U S A. 2012 Feb 21;109(8):2724-9. Iyer MK, Chinnaiyan AM, Maher CA. ChimeraScan: a tool for identifying chimeric transcription in sequencing data. Bioinformatics. 2011 Oct 15;27(20):2903-4.

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S.L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol 10, R25.

Li, J., Smyth, P., Flavin, R., Cahill, S., Denning, K., Aherne, S., Guenther, S.M., O'Leary, J.J., and Sheils, O. (2007). Comparison of miRNA expression patterns using total RNA extracted from matched samples of formalin-fixed paraffin-embedded (FFPE) cells and snap frozen cells. BMC Biotechnol 7, 36.

Medeiros F, Rigl CT, Anderson GG, Becker SH, Halling KC. Tissue handling for genome-wide expression analysis: a review of the issues, evidence, and opportunities. Arch Pathol Lab Med. 2007 Dec;131(12):1805-16. Review.

Meissner A, Mikkelsen TS, Gu H, Wernig M, Hanna J, Sivachenko A, Zhang X, Bernstein BE, Nusbaum C, Jaffe DB, Gnirke A, Jaenisch R, Lander ES.Genome-scale DNA methylation maps of pluripotent and differentiated cells. Nature. 2008 Aug 7;454(7205):766-70. Epub 2008 Jul 6.

Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods. 2008 Jul;5(7):621-8. Epub 2008 May 30.

Oliver PG, LoBuglio AF, Zhou T, Forero A, Kim H, Zinn KR, Zhai G, Li Y, Lee CH, Buchsbaum DJ. Effect of anti-DR5 and chemotherapy on basal-like breast cancer. Breast Cancer Res Treat. 2012 Jun;133(2):417-26.

Sandoval, J., Heyn, H.A., Moran, S., Serra-Musach, J., Pujana, M.A., Bibikova, M., and Esteller, M. (2011). Validation of a DNA methylation microarray for 450,000 CpG sites in the human genome. Epigenetics 6, 692-702.

Tomlins SA, Rhodes DR, Perner S, Dhanasekaran SM, Mehra R, Sun XW, Varambally S, Cao X, Tchinda J, Kuefer R, Lee C, Montie JE, Shah RB, Pienta KJ, Rubin MA, Chinnaiyan AM. Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer. Science. 2005 Oct 28;310(5748):644-8.

Xi, Y., Nakajima, G., Gavin, E., Morris, C.G., Kudo, K., Hayashi, K., and Ju, J. (2007). Systematic analysis of microRNA expression of RNA extracted from fresh frozen and formalin-fixed paraffin-embedded samples. RNA 13, 1668-1674.

Zhulidov PA, Bogdanova EA, Shcheglov AS, Vagner LL, Khaspekov GL, Kozhemyako VB, Matz MV, Meleshkevitch E, Moroz LL, Lukyanov SA, Shagin DA. Simple cDNA normalization using kamchatka crab duplex-specific nuclease. Nucleic Acids Res. 2004 Feb 18;32(3):e37.

Zou KH, Hall WJ, Shapiro DE.Smooth non-parametric receiver operating characteristic (ROC) curves for continuous diagnostic tests. Stat Med. 1997 Oct 15;16(19):2143-56.

Supporting Data

Figure 1: DNA methylation in 28 breast cancer cell lines. Unsupervised hierarchical clustering of the 10,000 CpGs with the most variable DNA methylation across breast cancer cell lines. Yellow: full methylation. Blue: no methylation. Breast cancer cell lines exhibit distinct patterns of DNA methylation that distinguish the luminal and basal subtypes of breast cancer.

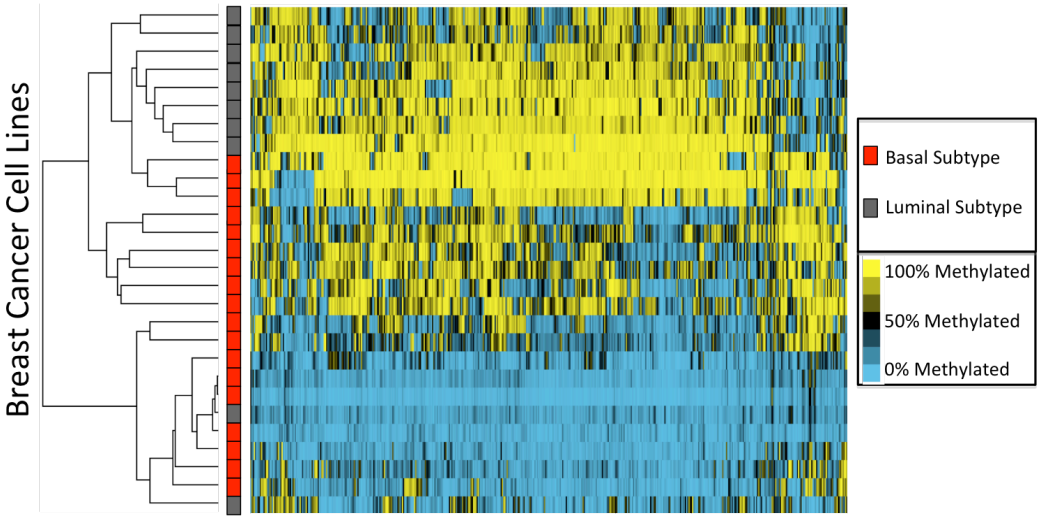


Figure 2: Principal Components Analysis (PCA) of DNA methylation patterns from 450,000 CpGs distinguishes prostate tumor (red) and benign-adjacent prostate (green) tissue samples.

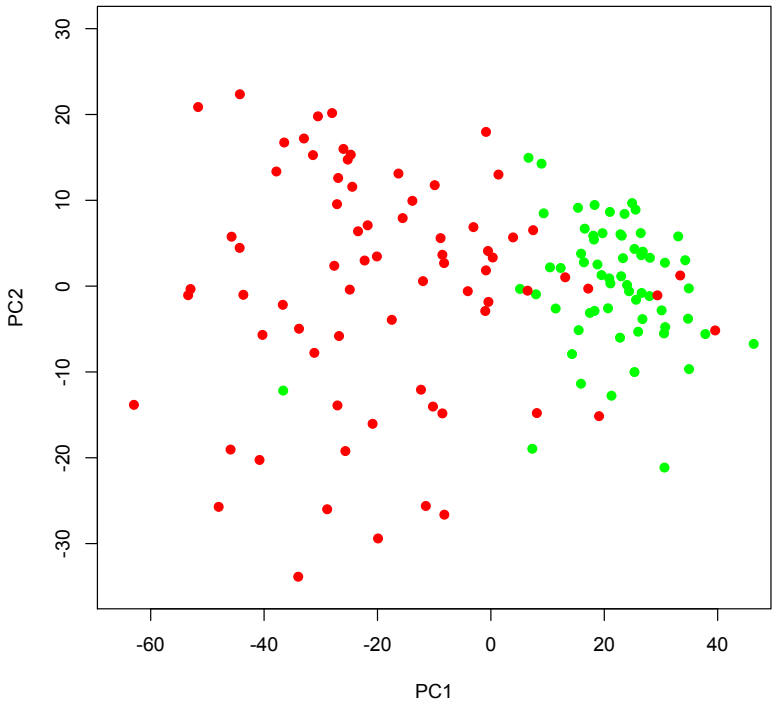


Figure 3: Unsupervised hierarchical clustering of ~12,000 most variable CpGs from prostate tumor tissue samples. Each column across the top is one individual's prostate tumor, and each of the 12,000 CpGs are on the vertical axis. Yellow: full methylation. Blue: represents no methylation.

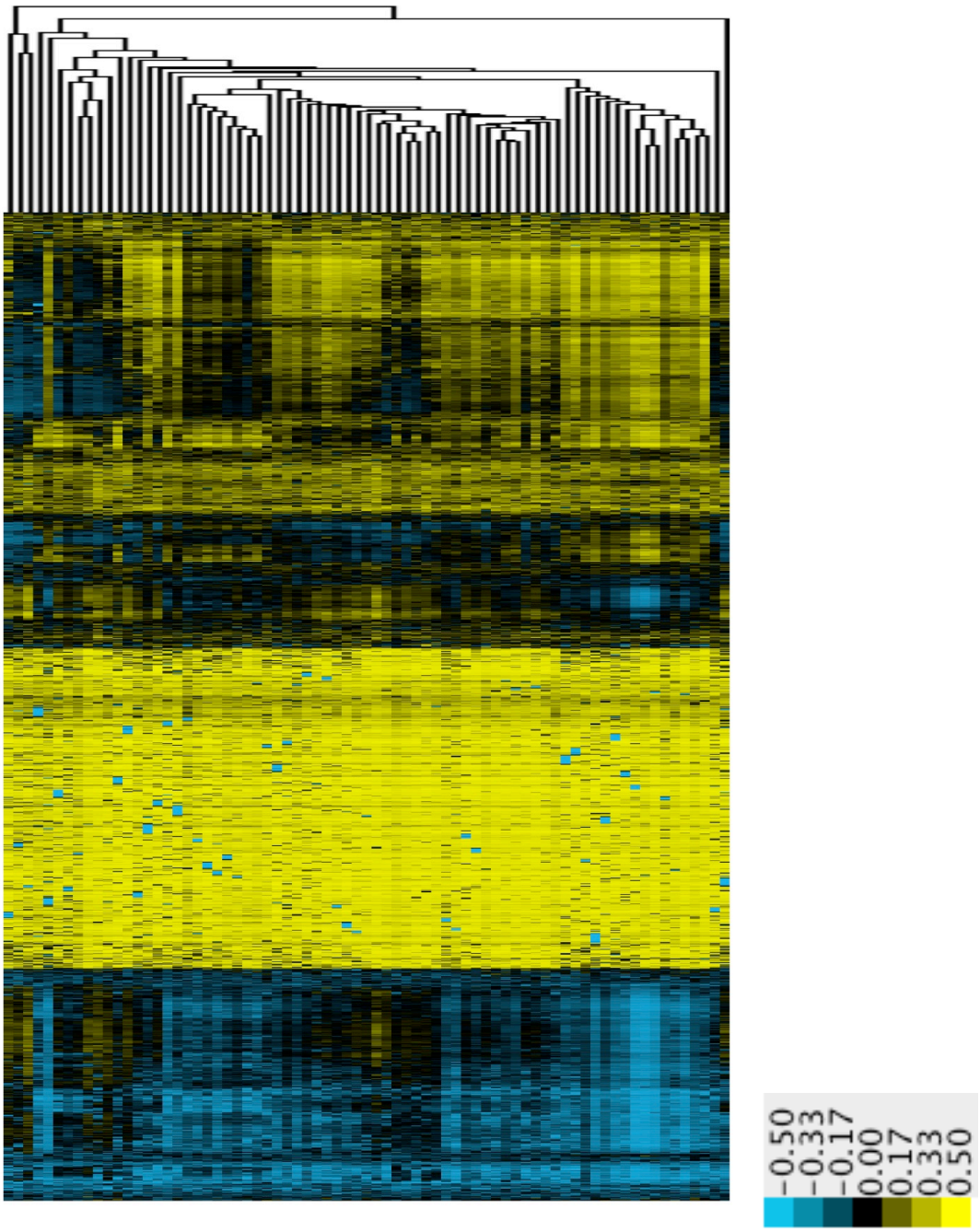
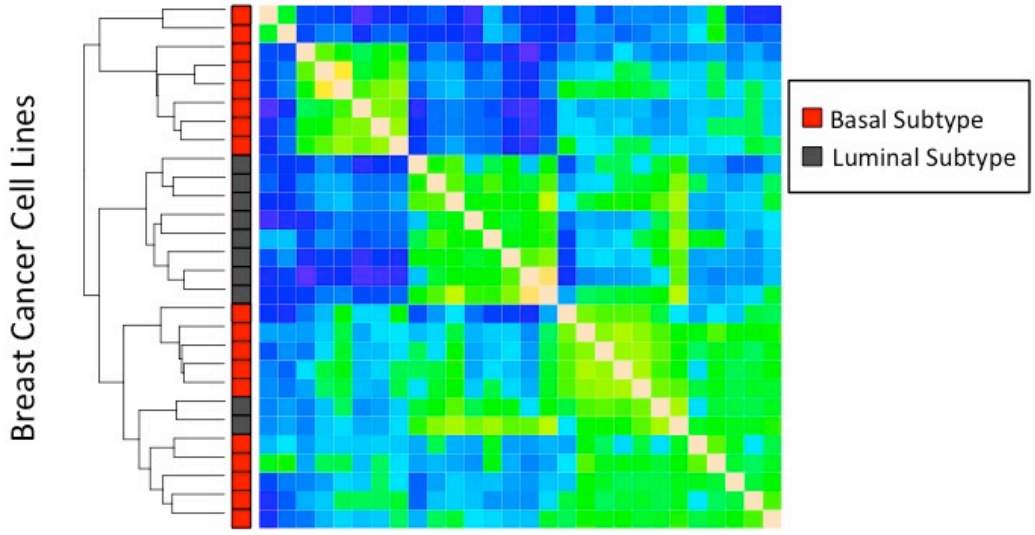


Figure 4: mRNA-seq in breast cancer. The correlation coefficient of the expression values between all pairs of breast cancer cell lines is displayed in this heatmap. Green and yellow represent highly correlated samples, and blue and cyan represent weakly correlated samples. The similarity of expression values corresponds to similar breast cancer subtypes. The dendogram on the left shows the similarity between the expression in the basal subtype samples and the expression signature shared by the luminal subtype samples.



List of personnel that received pay from this research effort:
Richard Myers
Greg Barsh
Devin Absher
Shawn Levy
Katherine Varley
Marie Cross
Tracy Eggleston
Barbara Pusey
Scott Newberry
Brittany Lasseigne
Stephanie Parker
Alan You
Tatsuya Uechi
Joshua Nielson
Todd Burwell
Preti Jain

Appendices

The following appendices are attached to this document in order:

Appendix A: A manuscript that describes the improved library protocol for RNA-seq experiments was accepted for publication in Genome Research

Appendix B: A manuscript describing the discovery of read-through fusion transcripts in breast cancer.

Appendix C: An abstract describing the DNA methylation data and preliminary analysis from the prostate samples submitted to the AGBT 2012 conference.

Transposase mediated construction of RNA-seq libraries

Jason Gertz, Katherine E. Varley, Nicholas S. Davis, et al.

Genome Res. 2012 22: 134-141 originally published online November 29, 2011
Access the most recent version at doi:10.1101/gr.127373.111

Method

Transposase mediated construction of RNA-seq libraries

Jason Gertz,[1] Katherine E. Varley,[1] Nicholas S. Davis,[1] Bradley J. Baas,[2] Igor Y. Goryshin,[2] Ramesh Vaidyanathan,[2] Scott Kuersten,[2] and Richard M. Myers[1,3]

[1]*HudsonAlpha Institute for Biotechnology, Huntsville, Alabama 35806, USA;* [2]*Epicentre (An Illumina Company), Madison, Wisconsin 53713, USA*

RNA-seq has been widely adopted as a gene-expression measurement tool due to the detail, resolution, and sensitivity of transcript characterization that the technique provides. Here we present two transposon-based methods that efficiently construct high-quality RNA-seq libraries. We first describe a method that creates RNA-seq libraries for Illumina sequencing from double-stranded cDNA with only two enzymatic reactions. We generated high-quality RNA-seq libraries from as little as 10 pg of mRNA (~1 ng of total RNA) with this approach. We also present a strand-specific RNA-seq library construction protocol that combines transposon-based library construction with uracil DNA glycosylase and endonuclease VIII to specifically degrade the second strand constructed during cDNA synthesis. The directional RNA-seq libraries maintain the same quality as the nondirectional libraries, while showing a high degree of strand specificity, such that 99.5% of reads map to the expected genomic strand. Each transposon-based library construction method performed well when compared with standard RNA-seq library construction methods with regard to complexity of the libraries, correlation between biological replicates, and the percentage of reads that align to the genome as well as exons. Our results show that high-quality RNA-seq libraries can be constructed efficiently and in an automatable fashion using transposition technology.

[Supplemental material is available for this article.]

RNA-seq is a powerful technique that allows for sensitive digital quantification of transcript levels (Mortazavi et al. 2008; Nagalakshmi et al. 2008). It enables the detection of noncanonical transcription start sites (Liu et al. 2011) as well as termination sites (Wang et al. 2008), alternative splice isoforms (Wang et al. 2008; Jiang and Wong 2009), transcript mutations/editing (Rosenberg et al. 2011), and allelic biases in transcript abundance (Pickrell et al. 2010). Methods that preserve the strand from which the transcript originated also allow for the identification of antisense transcription (He et al. 2008; Perkins et al. 2009), which can play a role in post-transcriptional regulation. Because of the power of RNA-seq and the prevalence of aberrant gene-expression patterns in many diseases, there is a growing need to construct libraries efficiently from low starting amounts of RNA in a high-throughput and reproducible fashion.

Ultra-high throughput, "next-generation" DNA sequencing library construction is a time-consuming process that typically has some sample loss at each step. A recent advance in library construction is the use of transposases to randomly integrate sequencing adapters into the DNA of interest (Adey et al. 2010). This approach creates sequencing-ready DNA libraries in a few steps with minimal hands-on time. The resulting libraries exhibit even coverage across the human genome when constructed from low amounts of genomic DNA (Adey et al. 2010). Transposon-based library construction has also been successfully applied to pyrosequencing of the RNA genomes of strains of simian hemorrhagic fever virus (Lauck et al. 2011). The success of transposon-based genomic library construction for genomic analyses suggests that it should be possible to use transposases to construct high-quality RNA-seq libraries.

Recently, several techniques developed for constructing RNA-seq libraries which maintain the transcript strand-of-origin were evaluated (Levin et al. 2010). Each protocol had varying levels of strand specificity, library complexity, and reproducibility. One of the overall best methods tested involved incorporating uracil into the second cDNA strand. The strand is subsequently degraded specifically by treatment with uracil DNA glycosylase and endonuclease VIII, which leaves only sequence reads that map to the strand-of-origin of each transcript (Parkhomchuk et al. 2009). The application of transposases to construct strand-specific RNA-seq libraries is an appealing approach for efficiently creating RNA-seq libraries with maximal information.

Here we describe the development of a transposon-based method for RNA-seq library construction, called Tn–RNA-seq. The method is fast and requires only two steps and two purifications after cDNA is made. The protocol is fully automatable and is compatible with robotics. We also extend and modify the transposase-based RNA-seq method to create directional RNA-seq libraries capable of preserving the strand information from which the transcript originated.

Results

Efficient transposition-based RNA-seq library construction

The strategies of each protocol are outlined in Figure 1. To construct nondirectional standard RNA-seq libraries, we prepared double-stranded (ds) cDNA from fragmented mRNA (Mortazavi et al. 2008). The ds cDNA was end-repaired, A-tailed, ligated to sequencing adapters, and amplified (Fig. 1A). For the nondirectional transposon-based RNA-seq method (Tn–RNA-seq), mRNA was not frag-

[3]**Corresponding author.**
E-mail rmyers@hudsonalpha.org.
Article published online before print. Article, supplemental material, and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.127373.111.

22:134–141 © 2012 by Cold Spring Harbor Laboratory Press; ISSN 1088-9051/12; www.genome.org

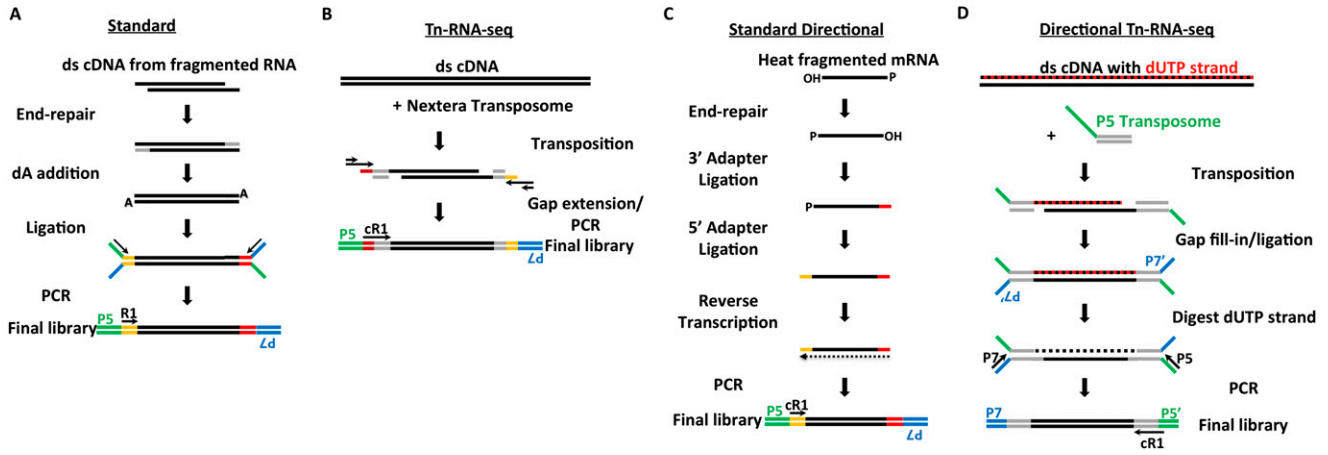


Figure 1. RNA-seq methods overview. (*A*) In the standard adapter ligation RNA-seq library construction protocol, double-stranded cDNA made from fragmented mRNA is subjected to end repair, dATP addition, adapter ligation, size selection, and PCR. (*B*) For the Tn–RNA-seq method described here, double-stranded cDNA is incubated with transposome (transposase complexed with transposon) and then undergoes PCR. (*C*) In the directional RNA ligation approach (standard directional), poly(A)-selected mRNA is fragmented with heat and end repaired. 3′ and then 5′ adapters are ligated onto single-stranded RNA before reverse transcription, followed by PCR. (*D*) The directional Tn–RNA-seq library construction described here starts with double-stranded cDNA, in which the second strand synthesized contains uracils instead of thymines. cDNA synthesis is followed by transposition of sequencing adapters, gap fill-in/ligation, USER digestion, and PCR. P5 and P7 correspond to Illumina cluster generation primers. R1 identifies the sequencing primer and cR1 indicates custom sequencing primer. (′) Reverse complement.

mented before cDNA synthesis. Instead, we incubated the ds cDNA with a transposome (hyperactive Tn5 transposase bound to synthetic 19-bp mosaic end-recognition sequences appended to Illumina sequencing adapters) (Adey et al. 2010) to simultaneously fragment and attach adapters (Fig. 1B). Because the transposome is a mixture containing two different sequences (shown in red and yellow in Fig. 1B), it can insert in either orientation, resulting in a nondirectional library. During the transposition process, only the transferred strand of each transposon end is covalently linked to the target DNA. Due to the staggered fashion of the transposition, a 9-bp gap between the nontransferred strand and the target DNA is created. Extension synthesis from the target DNA into this gap, followed by copying of the attached transposon end by strand displacement, creates the 3′ adapter sequence. Suppression PCR (Rand et al. 2005) is then used to select for templates with heterologous adapters. During PCR, index barcodes can be added to allow for the mixing of multiple samples in one sequencing lane. Following purification, PCR products are ready for single-end or paired-end sequencing with custom sequencing primers.

The entire process is automatable and feasible in 96-well plates, making large-scale Tn–RNA-seq library construction with robotics an appealing combination. Standard RNA-seq library construction (Mortazavi et al. 2008) requires multiple enzymatic reactions between cDNA synthesis and the final PCR step, compared with the one reaction with the Tn–RNA-seq method we describe here (Fig. 1B). The Tn–RNA-seq protocol cuts down significantly on sample preparation time and could yield higher quality RNA-seq libraries by minimizing sample loss during multiple reactions and purifications.

We constructed RNA-seq libraries with the standard method and the Tn–RNA-seq method to compare library quality. We extracted high-quality total RNA (RNA integrity number of 9.5 on an Agilent Bioanalyzer) from the human endometrial adenocarcinoma cell line ECC-1 (Mo et al. 2006), and purified ECC-1 mRNA by using poly(A) selection on magnetic beads. Two biological replicates were used for each method with a starting amount of 50 ng of ECC-1 mRNA. Each library was sequenced on one lane of an Illumina GAIIx using a custom sequencing primer specific to the transposon end to generate an average of 28 million pass-filter 36-bp reads. We used multiple metrics to assess quality of each library. The results are shown in Table 1 (see Methods for details on calculations) and some typical examples are presented in Figure 2 and Supplemental Figure S5. The libraries were evaluated on how well they aligned to exons, their complexity, and their biological reproducibility. In each of these categories, the Tn–RNA-seq protocol showed similar performance to the standard protocol, indicating that high-quality RNA-seq libraries can be constructed using transposition technology.

We also evaluated 5′ and 3′ bias in each library by calculating the relative coverage across transcripts. Figure 3 shows that the Tn–RNA-seq libraries show a subtle depletion in coverage at the 10%-most 5′ ends of transcripts. This depletion is due to the nature of the transposon-based library construction. To sequence the ends of transcripts, a transposase would have to integrate one transposon near the very end of the transcript and that transposon would have to be in the correct orientation with the sequencing primer facing toward the 3′ end of the transcript. Even in this case, only sequencing reads generated from one strand would map to the most 5′ end of the transcript. Depletion is not seen on the 3′ end of the transcript, which is most likely due to the presence of the poly(A) tail, which gives the transposase extra substrate to integrate the transposon. The depletion of 5′ ends is seen across all sizes of transcripts (Supplemental Fig. S2), and when analyzed on a base-pair scale, corresponds to a more than twofold depletion in coverage of the first 50 nt of the transcript relative to the standard method (Supplemental Fig. S3). We observed a subtle reduction in the number of short transcripts, <200 nt, which were detectable with the Tn–RNA-seq approach. We found that 396 short transcripts were detectable with the standard approach and 348 short transcripts were detectable with the Tn–RNA-seq method, which represents a 12% reduction in the number of short transcripts detected. It is important to point out that the higher alignment percentage to exons of the Tn–RNA-seq method, compared with the standard protocol, may be due to the depletion of 5′ ends,

Table 1. RNA-seq library construction comparison in ECC-1

Protocol	Reads aligned to genome	Percent of aligned reads that map to RefSeq transcripts	Complexity	Biological Replicate Correlation (Pearson)	Biological Replicate Correlation (Spearman)
Standard Rep1	18,347,613	73.7%	85.77%	0.983	0.986
Standard Rep2	9,218,462	74.3%	85.79%		
Tn–RNA-seq Rep1	22,945,616	81.3%	89.08%	0.955	0.986
Tn–RNA-seq Rep2	18,513,940	85.1%	85.20%		
Directional Tn–RNA-seq Rep1	20,474,907	74.5%	81.60%	0.981	0.988
Directional Tn–RNA-seq Rep2	25,848,260	72.5%	81.43%		

because observed 5′ ends of transcripts may differ from RefSeq annotations. While there is a slight depletion in coverage at the 5′ end of transcripts, the impact on library quality and expression measurements (discussed below) is negligible.

To determine whether the Tn–RNA-seq protocol produces data indicating the same gene-expression levels as does the standard protocol, we calculated RPKM (reads per kilobase per million aligned reads) (Mortazavi et al. 2008) values for each RefSeq gene (see Methods). The results are shown in Figure 4. The Pearson correlation (r) between log base 2 of RPKM values from the standard and Tn–RNA-seq protocols is 0.959. The Spearman rank correlation (ρ), which is more appropriate given the overall distribution of RPKM values, is 0.979. The high correlation in expression values indicates that the Tn–RNA-seq protocol allows for efficient construction of high-quality RNA-seq libraries while maintaining the integrity of transcript measurements.

Consistent Tn–RNA-seq libraries constructed from low amounts of input mRNA

We next sought to establish whether the Tn–RNA-seq method is robust to differing amounts of starting material and determine the amount of mRNA required to construct a reliable RNA-seq library. To test library construction with lower starting amounts of mRNA, we constructed seven Tn–RNA-seq libraries with between 10 ng and 1 pg of mRNA. The yield of Tn–RNA-seq library construction was dependent on the amount of mRNA that was used. Libraries made with between 10 and 0.5 ng of mRNA yielded ~600 ng of DNA, while libraries constructed with <100 pg of mRNA yielded between 30 and 100 ng of DNA.

The quality metrics for each Tn–RNA-seq library are displayed in Supplemental Table S1. All six libraries made from between 10 ng and 10 pg of mRNA had at least 72% of aligned reads map to known transcripts, while the library made from 1 pg of mRNA had 62% of aligned reads map to known transcripts. Library complexity also remained high for all libraries except for the library constructed with 1 pg of mRNA (Fig. 5A). In general, Tn–RNA-seq libraries made with 10 pg or more of mRNA exhibited consistent quality measures, showing that high-quality RNA-seq libraries can be constructed with the transposon-based method from as little as 10 pg of mRNA, which represents ~1 ng of total RNA or ~200 cell equivalents.

We also examined whether expression measurements were consistent across different amounts of starting materials (Supplemental Fig. S4). We found that all libraries made from at least 10 pg of mRNA were very consistent with the libraries constructed from 50 ng of mRNA. For all libraries except for the library made with 1 pg of mRNA, the rank correlation of expression measurements with the 50 ng of mRNA library exceeded 0.96 (Fig. 5B).

Library insert size is influenced by the amount of mRNA used; smaller amounts of starting material result in smaller insert sizes (Supplemental Fig. S1). This is due to the relative ratio of target DNA to transposome, since the transposase does not enzymatically turn over in these reactions. Based on these observations, it may be possible to alter the insert size by changing the concentration of transposase relative to the amount of mRNA. Our results indicate that transposon-based library construction can be used on limiting amounts of mRNA as low as 10 pg.

Strand-specific transposon-based RNA-seq library construction

A limitation of the above-mentioned transposon-based approach is that the transposition reaction is inherently nondirectional. This means that the resulting cDNA is captured without regard to the original transcript strand information. To create libraries that preserve strand information we adapted a previously described approach to specifically mark one strand of cDNA by incorporating

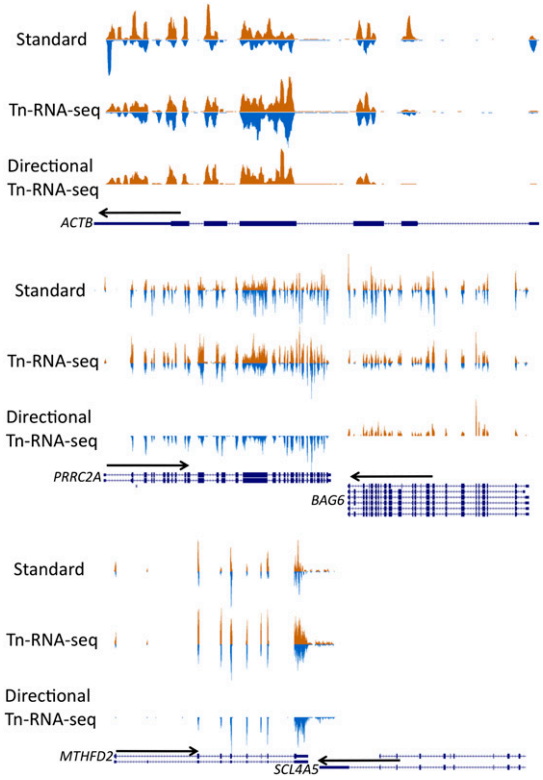


Figure 2. The directional Tn–RNA-seq method exhibits complete strand specificity. Aligned reads for the standard, Tn–RNA-seq, and directional Tn–RNA-seq method are displayed for three genomic loci. Reads mapping to the positive strand are shown in orange and reads mapping to the negative strand are shown in blue. Arrows indicate the direction of transcription for each RefSeq gene.

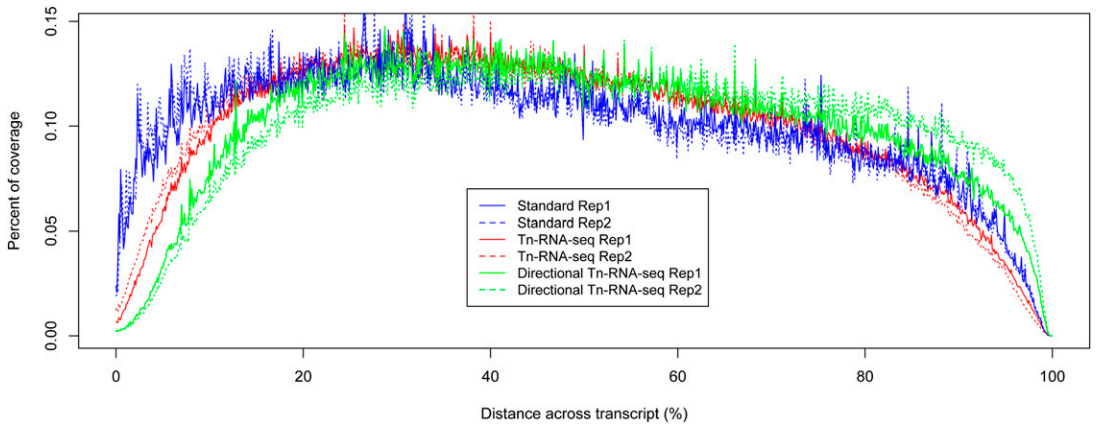


Figure 3. Transposon-based libraries show expected depletion of coverage at 5′ ends of transcripts. The percentage of coverage (*y*-axis), averaged across all transcripts is plotted as a function of distance across the transcripts (*x*-axis). 0% corresponds to the 5′ ends and 100% corresponds to the 3′ ends of transcripts.

dUTP during the second-strand synthesis (Parkhomchuk et al. 2009). We modified the Tn–RNA-seq method to accommodate uracil-containing cDNA and preserve the stranded information content of the samples. After adapters were attached, the combination of uracil DNA glycosylase (UDG) and endonuclease VIII (Endo VIII) degraded the second strand, leaving only the first strand of cDNA, which is the reverse complement of the original transcript.

After first-strand cDNA synthesis from 50 ng of ECC-1 mRNA, we treated the reaction with a nucleotide phosphatase to remove nucleotides, since free nucleotide contamination in the second-strand reaction would result in a decrease in strand specificity. The reaction was then column purified and used for second-strand cDNA synthesis in the presence of a nucleotide mix containing dUTP instead of dTTP.

Purified uracil-containing double-stranded cDNA was then incubated with a single transposome containing a unique sequence (P5), which is appended to the 5′ end of the transferred strand (Fig. 1D, shown in green). After transposition, DNA fragments contain the P5 sequence at the 5′ ends of both strands of cDNA. The nontransferred strand is replaced with a modified oligonucleotide containing a different sequence (P7) appended to the 5′ end (Fig. 1D, shown in blue) and the 9-bp gap is filled in and ligated to the template. Because the cDNA is marked, the uracil-containing second strand can be removed prior to PCR by treating the cDNA library with UDG and Endo VIII. The surviving fragments are then amplified and enriched using Phusion DNA polymerase, which is very inefficient at extending templates that contain uracils, providing an additional level of strand specificity (Greagg et al. 1999). We sequenced 36 bases of these final libraries from a single end on an Illumina GAIIx using a custom sequencing primer specific to the P5-containing transposon end. This directional library method (directional Tn–RNA-seq) is designed to produce all sequencing tags oriented 3′–5′ relative to the original RNA transcripts.

We observed striking strand specificity in the genomic alignments produced from these libraries (Fig. 2). *ACTB*, one of the highest expressed genes in ECC-1, is shown in Figure 2, top, and we found that all reads map to the expected strand. Determining strand specificity can help to disambiguate some genes; for example,

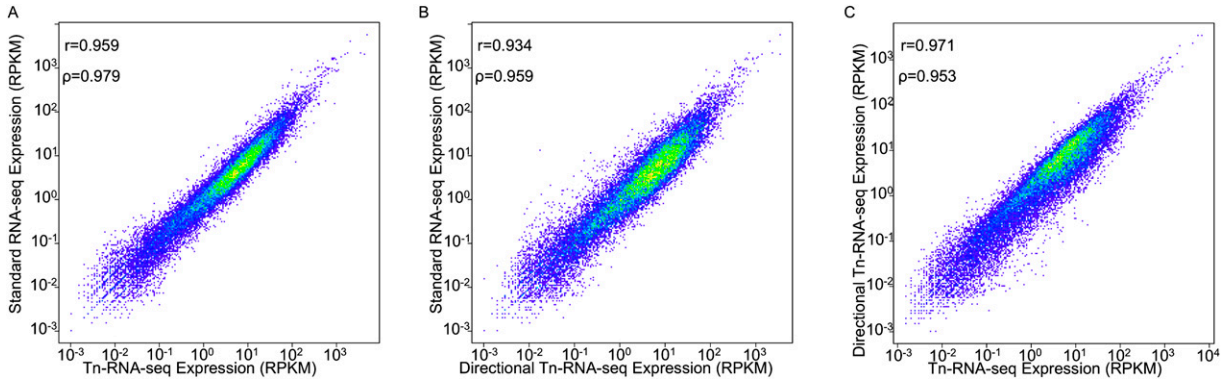


Figure 4. Expression values are consistent between standard RNA-seq library construction and transposon-based RNA-seq library construction in ECC-1. (*A*) Scatterplot showing expression values for standard RNA-seq library construction (*y*-axis) and the Tn–RNA-seq library construction (*x*-axis). The Pearson correlation between the standard and Tn–RNA-seq protocols is 0.959, and the Spearman rank correlation is 0.979. (*B*) Scatterplot displaying expression values for standard RNA-seq library construction (*y*-axis) and the directional Tn–RNA-seq library construction (*x*-axis). The Pearson correlation between the standard and directional Tn–RNA-seq protocols is 0.934, and the Spearman rank correlation is 0.959. (*C*) Scatterplot displaying expression values for Tn–RNA-seq library construction (*x*-axis) and the directional Tn–RNA-seq library construction (*y*-axis). The Pearson correlation between the Tn–RNA-seq and directional Tn–RNA-seq protocols is 0.971, and the Spearman rank correlation is 0.953.

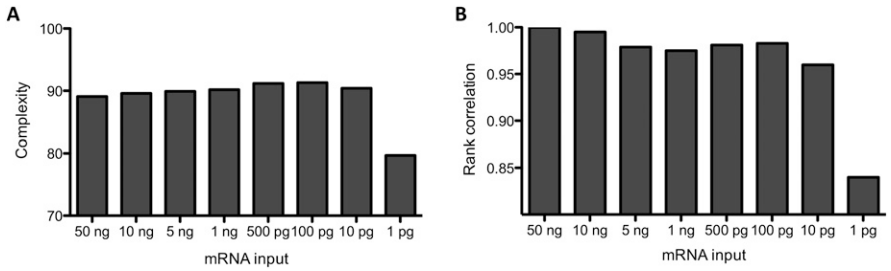


Figure 5. Tn–RNA-seq libraries constructed from as low as 10 pg of mRNA are high quality and show highly correlated expression levels. (*A*) Library complexity, calculated as the number of different alignment positions in a random set of 1 million aligned reads divided by 1 million, is shown for libraries made with between 50 ng and 1 pg of mRNA. (*B*) Rank correlations of expression measurements between the library constructed with 50 ng of mRNA and every other Tn–RNA-seq library are displayed.

SLC4A5 (Fig. 2, bottom). When measured by nondirectional RNA-seq, *SLC4A5* appears expressed because of reads mapping to the 3′ end of the gene. However, strand-specific RNA-seq shows that these reads originate from the antisense strand, not the coding strand, and represent either antisense transcription or read-through of the 3′ end of *MTHFD2*, a gene 3′ and oriented opposite to the *SLC4A5* gene.

To determine the overall strand specificity of the directional Tn–RNA-seq method, we calculated the percentage of reads mapping to the expected strand of RefSeq genes. We found that in both replicates >99.4% of reads map to the expected strand (99.5% and 99.4% for individual replicates). This is likely an underestimate of the strand specificity of the method, as there is expected to be some antisense transcription as well as alternative 5′ and 3′ UTR boundaries (as may be the case with *SLC4A5*) that are not represented in the RefSeq annotations. This level of strand specificity is in the same range as the most strand-specific methods analyzed by Levin et al. (2010) in yeast, indicating that the directional Tn–RNA-seq method exhibits a degree of strand specificity that is comparable to the most specific methods available.

The coverage across the length of transcripts for the directional Tn–RNA-seq RNA-seq libraries yields an interesting pattern. We observed substantial depletion at the 5′ end of transcripts and increased coverage at the 3′ end of transcripts compared with the standard RNA-seq library construction protocol. This pattern can be explained by the strand specificity of the directional Tn–RNA-seq method. Strand-specific reads in these libraries should always sequence from the 3′ end toward the 5′ end of the transcript. This would cause depletion in coverage at the 5′ end, because two transposition events near the 5′ end would be required to sequence the 5′−most portion of the transcript. The 3′ end harbors an overabundance of reads compared with the standard method because, regardless of where the transposon is integrated, the 3′ most transposon will be the sequencing primer that generates the sequence read.

While the directional Tn–RNA-seq method yields highly strand-specific libraries, we also wanted to assess the quality of the libraries using the same metrics discussed above. Table 1 shows that the directional Tn–RNA-seq libraries have similar levels of reproducibility, complexity, and alignability. The complexity of the directional Tn–RNA-seq libraries is lower compared with the standard and nondirectional Tn–RNA-seq libraries. We believe that this is due in part to a reduction in the number of possible alignment locations. The number of possible unique genome mapping locations, which includes the strand that the read matches, is cut by half in the directional Tn–RNA-seq libraries due to the strand specificity. Overall, these results show that our directional Tn–RNA-seq protocol results in high-quality strand-specific RNA-seq libraries that preserve transcript measurements (Fig. 4B,C).

To compare the directional Tn–RNA-seq with a standard directional RNA-seq protocol, we created strand-specific RNA-seq libraries using single-stranded RNA ligation (Lister et al. 2008), similar to the Illumina TruSeq small RNA protocol (Fig. 1C; see Methods). We created strand-specific libraries from universal human reference RNA (Novoradovskaya et al. 2004) using both methods to compare performance. For each library, we calculated the percentage of reads mapping to the expected strand of RefSeq genes. Similar strand specificity was observed with each protocol. The library constructed with the standard directional approach exhibited 99.46% of reads mapping to the expected strand, and the library constructed with the directional Tn–RNA-seq method had 99.51% of reads mapping to the expected strand. We next analyzed expression measurements from the directional libraries and found a high correlation (rank correlation: 0.96) between the two methods (Supplemental Fig. S6). These results indicate that the directional Tn–RNA-seq method maintains the same strand specificity of a standard method, and that expression measurements are also consistent between directional approaches.

Discussion

We have described two techniques for constructing RNA-seq libraries that are based on the introduction of sequencing adapters by transposition into double-stranded cDNA. The first method described is an efficient method for creating strand ambiguous libraries that requires only one enzymatic step to go from double-stranded cDNA to fragments ready to be amplified before sequencing. We found that libraries constructed in this manner performed as well as libraries made using the more laborious standard adapter ligation-based approach. We also found that that the transposon-based approach yielded high-quality RNA-seq libraries that preserved transcript measurements with as low as 10 pg of mRNA. This reduction in required starting material for RNA-seq library construction provides the opportunity to create reproducible RNA-seq libraries from rare cell types or small samples. The transposon-based RNA-seq approach is an attractive option for RNA-seq library construction because of the protocol's efficiency and efficacy. This is especially true for labs preparing a large number of samples, with or without robotics, because the library preparation starting from poly(A) selection takes <8 h to complete.

We also present a method that preserves the strand-of-origin for each transcript sequenced. Knowing the strand orientation of the transcripts can lead to interesting findings about transcript structure (Core et al. 2008; He et al. 2008; Seila et al. 2008). The strand specificity of the directional Tn–RNA-seq method comes from specific digestion of the second cDNA strand combined with novel transposome modifications to control the attachment of specific sequences to the template cDNA. The directional method is more time consuming than the nondirectional transposon-based method, but it provides additional information while maintaining a high level of complexity and reproducibility. The strand specificity is near complete at 99.5% and similar to the best-

published methods in yeast (Levin et al. 2010) and to our results using an RNA ligation approach.

While both transposon-based RNA-seq library construction techniques exhibit high-quality sequencing results, there is a subtle depletion of sequence near the 5′ ends of transcripts that is more pronounced with the directional method. This depletion is expected and unavoidable for directional Tn–RNA-seq due to the nature of the transposition and strand specificity. This depletion at the 5′ ends of transcripts could be lessened by modifying the protocol to sequence toward the 3′ end of the transcript as opposed to toward the 5′ end of the transcript in the protocol presented.

The nondirectional Tn–RNA-seq library construction is amenable to large-scale library construction and automation. Because every enzymatic step is followed by magnetic bead purification (Hawkins et al. 1994), the full library construction protocol can easily be applied to a 96-well plate format, where steps can be completed with robotics. The protocol also allows for multiplex sequencing of samples (Smith et al. 2010). Molecular barcodes can be added during the final PCR step by using different primers, which can result in a significant cost savings. Since the transposase binds to a particular sequence, the sequencing adapters introduced are different from the standard Illumina adapters. Therefore, the Tn–RNA-seq libraries need to be mixed with a custom primer to be sequenced, but otherwise require no special experimental or computational accommodations. Both transposase-based approaches to constructing RNA-seq libraries that are described in this work provide an efficient and streamlined workflow to achieve high-quality characterization of the transcriptome comparable to the current more laborious methods.

Methods

Cell culture and mRNA isolation

The human endometrial cancer cell line ECC-1 was grown in RPMI-1640 (Invitrogen) supplemented with 10% fetal bovine serum (Hyclone) and 1% penicillin-streptomycin (Invitrogen). Two separate growth replicates were used to assess biological replication. To isolate total RNA, we used the Animal Tissue RNA Isolation kit (Norgen) with ~5 million cells scraped from a 100-mm cell culture dish. The samples are DNase treated during the purification, which is important because genomic DNA contamination can be efficiently made into sequencer-ready molecules during the transposition step. Universal human reference RNA was purchased from Agilent. After total RNA was purified, mRNA was enriched using the Dynabeads mRNA Purificaiton Kit (Invitrogen) with the following modifications. The beads were washed twice, instead of once, with Wash Buffer B before each elution. Each sample went through two rounds of binding, washing, and elution. Samples were eluted in 20 μL of Tris-HCL elution buffer during the final elution. All RNA and DNA concentrations were measured with a Qubit fluorometer (Invitrogen).

Standard RNA-seq library construction

Standard library construction was performed as previously described (Mortazavi et al. 2008). For each biological replicate, 50 ng of mRNA was used for each library.

Tn–RNA-seq library construction

Primer and adapter sequences for both transposon-based protocols can be found in Supplemental Table S2. To make cDNA, 1 μL (3 μg) of random hexamers (Invitrogen) was added to poly(A)-selected mRNA in a volume of 20 μL of Tris-HCL elution buffer and incubated at 65°C for 5 min, then placed on ice. First-strand cDNA synthesis was performed by adding 4 μL of First Strand Buffer (Invitrogen), 2 μL of 100 mM DTT (Invitrogen), 0.5 μL of RNaseOUT (Invitrogen), and 1 μL of Superscript II (200 U/μL, Invitrogen), and incubating the mix at 25°C for 12 min, 42°C for 50 min, then 70°C for 15 min. The second strand of cDNA was filled-in by adding 16 μL of water, 5 μL of 10× second Strand Buffer (500 mM Tris-HCl at pH 7.8, 50 mM MgCl2, 10 mM DTT), 3 μL of 10 mM dNTPs (New England Biolabs—NEB), 1 μL of RNase H (10 U/μL, Invitrogen), and 5 μL of DNA Polymerase I (10 U/μL, Invitrogen) to the first-strand reaction on ice, and then incubating at 16°C for 2.5 h. The cDNA was purified with AMPure beads (Beckman Coulter) according to the manufacturer's instructions and eluted in 15 μL of EB (Qiagen).

To incorporate sequencing adapters, we combined the purified cDNA with 4 μL of TA buffer (33 mM Tris-acetate at pH 7.5, 66 mM potassium acetate, 10 mM magnesium acetate, and 0.5 mM DTT) and 0.2 μL of Nextera Enzyme (Epicentre) on ice and incubated at 55°C for 5 min, and then placed the sample on ice. We added 30 μL of QG buffer (Qiagen) to stop the transposase reaction and purified the DNA with 90 μL of AMPure beads, eluting in 22 μL of EB. To PCR amplify the fragments, we added 25 μL of Nextera PCR buffer, 1 μL of 50× Nextera Primer Cocktail, 1 μL of Nextera Adapter 2, and 1 μL of Nextera PCR enzyme (Epicentre) to the purified fragments for a total volume of 50 μL. The reaction was incubated at 72°C for 3 min, then at 95°C for 30 sec, followed by 15 cycles of 95°C for 10 sec, 62°C for 30 sec, and 72°C for 3 min. We purified the PCR amplicons with 90 μL of AMPure beads per the manufacturer's instructions and eluted in 32 μL of EB. We sequenced libraries at a concentration of 6 pM on an illumina GAIIx sequencer with a custom sequencing primer designed to anneal to the transposon sequence. Libraries constructed with <10 ng of mRNA were barcoded and sequenced on an Illumina HiSeq 2000 with a custom sequencing primer and custom index primer. For optimal sequencing results, we found that using 75 ng or less of double-stranded cDNA in the transposition reaction is important. Using larger amounts of cDNA can lead to insert sizes of >1000 bp that do not sequence well.

Directional Tn–RNA-seq library construction

To construct the first strand of cDNA, 50 ng of mRNA in a volume of 20 μL were added to 1 μL (3 μg) of random hexamers (Invitrogen), heated to 65°C for 5 min, and placed directly on ice. Then, 6 μL of 5x first strand buffer, 1 μL of 100 mM DTT, 1 μL of 10 mM dNTPs [NEB], 0.5 μL of RNaseOUT [Invitrogen], 0.5 μL of Actinomycin D [120 ng/μL, Sigma], and 1 μL of Superscript III [200 U/μL, Invitrogen] were added to mRNA/primer mix at room temperature. The reaction was then incubated at 40°C for 90 min and heat inactivated at 70°C for 15 min. The reaction was cooled to 37°C and 1 μL of RNase H (10 U/μL, Invitrogen) and 1 μL of NTPhos (20 U/μL, Epicentre) were added. The reaction was incubated at 37°C for 30 min, then heat inactivated at 75°C for 15 min. The first-strand cDNA was purified with the QIAquick PCR Purification kit (Qiagen) per the manufacturer's instructions and eluted in 25 μL of EB (Qiagen). It was then purified further with a prepared G50 Sephadex column (USA Scientific) to ensure removal of unincorporated dNTPs.

The second strand of cDNA was created by mixing 25 μL of the previously purified single-stranded cDNA, 13 μL of water, 5 μL of NEBuffer 2, 2 μL of 25 mM dNTPs (with dUTP instead of dTTP), 1 μL of random hexamers, and 4 μL of Klenow exo- (5 U/μL, NEB). The reaction was incubated at 37°C for 30 min. The second-strand synthesis product was purified using the MinElute PCR

Purification kit (Qiagen) per the manufacturer's instructions and eluted in 15 μL of EB.

To add sequencing adapters to the double-stranded cDNA, 15 μL of cDNA product, 4 μL of TA buffer (33 mM Tris-acetate at pH 7.5, 66 mM potassium acetate, 10 mM magnesium acetate, and 0.5 mM DTT), and 1 μL of directional Tn–RNA-seq Enzyme mix (Epicentre, beta test material) were mixed together on ice. They were gently vortexed and incubated at 55°C for 5 min. A 30-μL aliquot of QG buffer (Qiagen) was added immediately after the reaction finished. The reaction was cleaned up with 90 μL of AMPure Beads according to the manufacturer's instructions (Beckman Coulter) and eluted in 11 μL of EB. The 11 μL of purified DNA was mixed with 4 μL of Replacement Oligo (Epicentre, beta test material) and 4 μl of Fill-in Reaction Buffer (Epicentre, beta test material) and incubated at 45°C for 1 min, then 37°C for 30 min. A 1-μL aliquot of Gap Filling Enzyme (Epicentre, beta test material) was added and incubated at 37°C for an additional 30 min. The reaction was cleaned up with 36 μL of AMPure beads according to the manufacturer's instructions and eluted in 26 μL of EB.

To digest the second strand of cDNA, the 26-μL purified DNA was added to 3 μL of T4 DNA Ligase buffer (NEB) and 1 μL of USER enzyme mix (1 U/μL, NEB). The reaction was incubated at 37°C for 30 min and the cDNA was purified with 54 μL of AMPure beads according to the manufacturer's instructions and eluted in 25 μL of EB. To amplify the fragments, 25 μL of USER-treated DNA was added to 1 μL of directional Tn–RNA-seq PCR Primer 1 and 1 μL of directional Tn–RNA-seq Primer 2 (Epicentre, beta test material) and 27 μL of Phusion PCR mix (NEB). The mix was incubated at 95°C for 2 min, then 18 cycles of 94°C for 10 sec, 60°C for 30 sec, and 72°C for 3 min were performed. The PCR amplicons were purified with 97 μL of AMPure beads according to the manufacturer's instructions and eluted in 32 μL of EB. Libraries were sequenced on an Illumina GAIIx at a concentration of 6 pM.

Standard directional RNA-seq library construction

Poly(A)-selected Universal Human Reference RNA (Agilent) was heated for 8 min at 94°C in 1x fragmentation buffer (40 mM Tris-Acetate at pH 8.1, 100 mM KOAc, 30 mM MgOAc) and purified using an RNeasy MinElute Kit (Qiagen). The purified and fragmented RNA was treated with 1 μL of Antarctic Phosphatase (5U/μL; NEB) for 30 min at 37°C, and then heat killed for 5 min at 65°C. The samples were then incubated with 2 μL of T4 Polynucleotide Kinase (2U/μL; Epicentre) and 0.7 mM ATP for 30 min at 37°C, followed by purification using an RNeasy MinElute kit (Qiagen). The end-repaired RNA was then ligated and amplified using the Illumina TruSeq small RNA kit. This involves sequential ligation of 3′ and 5′ adapters, followed by reverse transcription and PCR to amplify the completed libraries. Material of the expected size was purified from free adapter product using AMPure Beads.

Data analysis

Every sequence library was trimmed to only analyze the first 36 bases in order to make the results comparable. To assess library quality, the sequence reads from each library were aligned to the GRCh37/hg19 build of the human genome using bowtie (Langmead et al. 2009) with the –m 1 option, to guarantee unique mapping. Complexity was measured for each library by taking a random set of 1 million aligned reads and determining how many different alignment start positions (including strand information) were represented. The number of different alignment start positions was then divided by 1 million to calculate complexity. The percentage of aligned reads that map to RefSeq transcripts was determined by comparing bowtie alignments against the human genome to RefSeq transcript coordinates. The strand specificity of directional Tn–RNA-seq libraries was calculated by determining what percentage of reads that aligned to RefSeq transcripts map to the expected strand in the RefSeq annotation. Note that based on the library construction strategy, transcripts originating from the positive strand should generate sequencing reads that map to the negative strand with the directional Tn–RNA-seq method. In the standard directional approach, transcripts originating from the positive strand should generate reads that map to the positive strand.

To calculate expression levels in each library, sequence reads were aligned to a sequence database of all spliced RefSeq transcripts using Bowtie (Langmead et al. 2009) with the following parameters: -n 2 -k 1 -m 10, which allow reads to align to multiple transcripts in order to capture different isoforms. The number of reads aligning to each transcript was multiplied by 1 million, then divided by the length of the transcript in kilobases times the total number of aligned reads to calculate RPKM values. All correlation analysis was performed in R. All Pearson correlations were measured between log base 2 of RPKM values. Coverage across transcripts was calculated by counting the number of reads that align at each position in a RefSeq transcript and dividing the position by the number of base pairs of the full-length transcript.

Data access

The sequencing data and expression measurements from this study have been submitted to the NCBI Gene Expression Omnibus (GEO) (http://www.ncbi.nlm.nih.gov/geo/) under accession no. GSE32307.

Competing interest statement

B.J.B., I.Y.G., R.V., and S.K. are employees of Epicentre (An Illumina Company).

Acknowledgments

We thank Mark Maffitt, Barbara Wold and members of her lab, as well as members of the Myers lab for valuable discussions and contributions. Portions of this work were funded by USAMRMC/TATRC Contract W81XWH-10-1-0790 (to R.M.M.) and by NHGRI ENCODE Grant 5U54HG004576 (to R.M.M.).

References

Adey A, Morrison HG, Asan , Xun X, Kitzman JO, Turner EH, Stackhouse B, MacKenzie AP, Caruccio NC, Zhang X, et al. 2010. Rapid, low-input, low-bias construction of shotgun fragment libraries by high-density in vitro transposition. *Genome Biol* **11:** R119. doi: 10.1186/gb-2010-11-12-r119.

Core LJ, Waterfall JJ, Lis JT. 2008. Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. *Science* **322:** 1845–1848.

Greagg MA, Fogg MJ, Panayotou G, Evans SJ, Connolly BA, Pearl LH. 1999. A read-ahead function in archaeal DNA polymerases detects promutagenic template-strand uracil. *Proc Natl Acad Sci* **96:** 9045–9050.

Hawkins TL, O'Connor-Morin T, Roy A, Santillan C. 1994. DNA purification and isolation using a solid-phase. *Nucleic Acids Res* **22:** 4543–4544.

He Y, Vogelstein B, Velculescu VE, Papadopoulos N, Kinzler KW. 2008. The antisense transcriptomes of human cells. *Science* **322:** 1855–1857.

Jiang H, Wong WH. 2009. Statistical inferences for isoform expression in RNA-Seq. *Bioinformatics* **25:** 1026–1032.

Langmead B, Trapnell C, Pop M, Salzberg SL. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* **10:** R25. doi: 10.1186/gb-2009-10-3-r25.

Lauck M, Hyeroba D, Tumukunde A, Weny G, Lank SM, Chapman CA, O'Connor DH, Friedrich TC, Goldberg TL. 2011. Novel, divergent simian hemorrhagic Fever viruses in a wild ugandan red colobus

monkey discovered using direct pyrosequencing. *PLoS ONE* **6:** e19056. doi: 10.1371/journal.pone.0019056.
Levin JZ, Yassour M, Adiconis X, Nusbaum C, Thompson DA, Friedman N, Gnirke A, Regev A. 2010. Comprehensive comparative analysis of strand-specific RNA sequencing methods. *Nat Methods* **7:** 709–715.
Lister R, O'Malley RC, Tonti-Filippini J, Gregory BD, Berry CC, Millar AH, Ecker JR. 2008. Highly integrated single-base resolution maps of the epigenome in *Arabidopsis*. *Cell* **133:** 523–536.
Liu Y, Han D, Han Y, Yan Z, Xie B, Li J, Qiao N, Hu H, Khaitovich P, Gao Y, et al. 2011. Ab initio identification of transcription start sites in the Rhesus macaque genome by histone modification and RNA-Seq. *Nucleic Acids Res* **39:** 1408–1418.
Mo B, Vendrov AE, Palomino WA, DuPont BR, Apparao KB, Lessey BA. 2006. ECC-1 cells: a well-differentiated steroid-responsive endometrial cell line with characteristics of luminal epithelium. *Biol Reprod* **75:** 387–394.
Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. 2008. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods* **5:** 621–628.
Nagalakshmi U, Wang Z, Waern K, Shou C, Raha D, Gerstein M, Snyder M. 2008. The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science* **320:** 1344–1349.
Novoradovskaya N, Whitfield ML, Basehore LS, Novoradovsky A, Pesich R, Usary J, Karaca M, Wong WK, Aprelikova O, Fero M, et al. 2004. Universal Reference RNA as a standard for microarray experiments. *BMC Genomics* **5:** 20. doi: 10.1186/1471-2164-5-20.
Parkhomchuk D, Borodina T, Amstislavskiy V, Banaru M, Hallen L, Krobitsch S, Lehrach H, Soldatov A. 2009. Transcriptome analysis by strand-specific sequencing of complementary DNA. *Nucleic Acids Res* **37:** e123. doi: 10.1093/nar/gkp596.
Perkins TT, Kingsley RA, Fookes MC, Gardner PP, James KD, Yu L, Assefa SA, He M, Croucher NJ, Pickard DJ, et al. 2009. A strand-specific RNA-Seq analysis of the transcriptome of the typhoid bacillus *Salmonella typhi*. *PLoS Genet* **5:** e1000569. doi: 10.1371/journal.pgen.1000569.
Pickrell JK, Marioni JC, Pai AA, Degner JF, Engelhardt BE, Nkadori E, Veyrieras JB, Stephens M, Gilad Y, Pritchard JK. 2010. Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature* **464:** 768–772.
Rand KN, Ho T, Qu W, Mitchell SM, White R, Clark SJ, Molloy PL. 2005. Headloop suppression PCR and its application to selective amplification of methylated DNA sequences. *Nucleic Acids Res* **33:** e127. doi: 10.1093/nar/gni120.
Rosenberg BR, Dewell S, Papavasiliou FN. 2011. Identifying mRNA editing deaminase targets by RNA-Seq. *Methods Mol Biol* **718:** 103–119.
Seila AC, Calabrese JM, Levine SS, Yeo GW, Rahl PB, Flynn RA, Young RA, Sharp PA. 2008. Divergent transcription from active promoters. *Science* **322:** 1849–1851.
Smith AM, Heisler LE, St Onge RP, Farias-Hesson E, Wallace IM, Bodeau J, Harris AN, Perry KM, Giaever G, Pourmand N, et al. 2010. Highly-multiplexed barcode sequencing: an efficient method for parallel analysis of pooled samples. *Nucleic Acids Res* **38:** e142. doi: 10.1093/nar/gkg368.
Wang ET, Sandberg R, Luo S, Khrebtukova I, Zhang L, Mayr C, Kingsmore SF, Schroth GP, Burge CB. 2008. Alternative isoform regulation in human tissue transcriptomes. *Nature* **456:** 470–476.

Received June 21, 2011; accepted in revised form October 17, 2011.

Errata

***Genome Research* 22:** 134–141 (2012)

Transposase mediated construction of RNA-seq libraries
Jason Gertz, Katherine E. Varley, Nicholas S. Davis, Bradley J. Baas, Igor Y. Goryshin, Ramesh Vaidyanathan, Scott Kuersten, and Richard M. Myers

On page 139 of the above-mentioned article, the sentence beginning on the second line of the second column is missing data and should read as follows:

First-strand cDNA synthesis was performed by adding 4 μL of First Strand Buffer (Invitrogen), 2 μL of 100 mM DTT (Invitrogen), 0.5 μL of RNaseOUT (Invitrogen), 1 μL of 10 mM dNTPs (NEB), and 1 μL of Superscript II (200 U/μL, Invitrogen), and incubating the mix at 25°C for 12 min, 42°C for 50 min, then 70°C for 15 min.

The authors apologize for any confusion this may have caused.

***Genome Research* 19:** 1429–1440 (2009)

Clusters and superclusters of phased small RNAs in the developing inflorescence of rice
Cameron Johnson, Anna Kasprzewska, Kristin Tennessen, John Fernandes, Guo-Ling Nan, Virginia Walbot, Venkatesan Sundaresan, Vicki Vance, and Lewis H. Bowman

The 22-nt small RNA that is predicted to set phasing of the phased 24-mer clusters was incorrectly denoted as miR2775 in two places in the paper (pages 1436 and 1437) and in two places in the Supplemental Material (pages 1 and 60, Supp. Fig. 6 and legend). The correct name for this miRNA is miR2275.

The authors apologize for any confusion this may have caused.

22:592 © 2012 by Cold Spring Harbor Laboratory Press; ISSN 1088-9051/12; www.genome.org

Recurrent read-through fusion transcripts in breast cancer

Katherine E. Varley[1], Jason Gertz[1], Kevin M. Bowling[1], Marie K. Cross[1], Nicholas S. Davis[1#], Amy S. Nesmith[1], Patsy G. Oliver[2], Brian S. Roberts[1], William E. Grizzle[2], Andres Forero-Torres[2], Donald J. Buchsbaum[2], Albert F. LoBuglio[2], Richard M. Myers[1]*.

[1]HudsonAlpha Institute for Biotechnology, Huntsville, AL, USA.

[#]Current affiliation: Duke University Graduate Program in Genetics and Genomics, Durham, NC, USA.

[2]University of Alabama at Birmingham Comprehensive Cancer Center, Birmingham, AL, USA.

* To whom correspondence should be addressed:
Richard M. Myers, Ph.D.
HudsonAlpha Institute for Biotechnology
601 Genome Way
Huntsville, AL 35806
Telephone: 256-327-0431
FAX: 256-327-0978
rmyers@hudsonalpha.org

Abstract:

Recurrent chromosomal rearrangements that create fusion genes with oncogenic activity represent powerful biomarkers and drug targets, classically in hematologic cancers and sarcomas, and more recently in prostate, lung and breast cancer. As improved technologies enable whole transcriptome sequencing, fusion transcripts created by splicing mRNAs from two different genes have been identified in the absence of DNA rearrangement. Read-through fusion transcripts that result from the splicing of two adjacent genes in the same coding orientation are particularly prevalent chimeric RNAs, and specific fusion transcripts have recently been associated with cellular proliferation and disease progression in prostate cancer. Here we report the discovery of read-through fusion transcripts in breast cancer using paired-end RNA sequencing (RNA-seq) of 168 breast samples, including breast cancer cell lines, triple negative breast cancer primary tumors, estrogen receptor positive breast cancer primary tumors, non-neoplastic breast tissue, and a collection of other normal human tissue controls. We identified three recurrent read-through fusion transcripts that are associated with breast cancer (*IL17RC-CRELD1*, *SCNN1A-TNFRSF1A,* and *CTSD-IFITM10*), and Western blots indicated they are translated into fusion proteins in breast cancer cells. Read-through fusion transcripts between adjacent genes with different biochemical functions represent a new type of recurrent molecular defect in breast cancer that warrant further investigation as potential biomarkers and therapeutic targets.

Introduction:

Fusion genes with oncogenic activity were first identified in hematologic malignancies, where chromosomal translocations frequently join two genes that result in an aberrant protein product (1, 2). These fused genes have been valuable prognostic markers and therapeutic targets (3). The therapeutic value of identifying fusion genes is exemplified by the development of selective inhibitors targeted to the ABL kinase involved in the BCR-ABL fusion that is present in 95% of patients with chronic myelogenous leukemia (1, 2, 4). Most recurrent fusion genes have been identified in leukemias, lymphomas, and soft tissue sarcomas where cytogenetic approaches to detect chromosomal aberrations using spectral karyotyping, fluorescent in situ hybridization, and flow cytometry have been developed (5). Cytogenetic approaches to detect fusion genes in the more common forms of cancer, epithelial tumors, are hampered by the poor chromosome morphology, complex karyotypes, and cellular heterogeneity that typify these tumors, although it has been posited that fusion genes are likely drivers of oncogenesis in these tumors as well (3, 5, 6). Until recently, the most prevalent recurrent fusion genes identified in breast cancer were the ETV6-NTRK3 fusion in secretory breast carcinoma, a rare subtype of infiltrating ductal carcinoma (7) and the MYB-NFIB fusion in adenoid cystic carcinomas, another rare form of breast cancer (8). Recently, genome-wide microarray profiling, whole genome sequencing and whole transcriptome

sequencing have made it possible to systematically identify fusion genes in solid tumors. With these methods, recurrent fusions that contribute to malignancy have been identified in prostate cancer (e.g. TMPRSS2 fused to ETS family transcription factors (9-11)), in lung cancer (EML4-ALK (12)), and in breast cancer (MAST kinases fused to NOTCH family genes (13)). New technologies and informatics approaches are enabling the identification of recurrent fusion genes in more common epithelial cancers that may serve as valuable biomarkers and drug targets (13-19).

In addition to fusion genes created by genomic rearrangements, fusion transcripts created by *cis*- and *trans*-splicing of mRNA, in the absence of a DNA rearrangements, have been detected by sequencing cDNA clone libraries and performing RNA-seq (20). These chimeric RNAs have been detected at low levels in expressed sequence tag (EST) libraries (21-23) and low levels across benign and malignant samples (6, 20, 24). One particularly prevalent class of chimeric RNAs involves adjacent genes in the same coding orientation that are spliced together to form an in-frame chimeric transcript that spans both genes. In recent literature, these have been referred to as read-through gene fusions, transcription-induced chimeras, co-transcription of adjacent genes coupled with intergenic splicing (CoTIS), or conjoined genes. Several of these read-through fusion transcripts have been identified specifically in prostate cancer and are associated with cellular proliferation and disease progression (25-33). Recurrent read-through transcripts have not yet been characterized in breast cancer. We used paired-end RNA-seq to identify recurrent read-through gene fusions in breast samples, and determined that three recurrent read-through fusion transcripts were associated with breast cancers. We quantified the percentage of transcript that is involved in each fusion and demonstrated the presence of protein products for these fused genes using Western blots.

Results and Discussion:

While recent studies have reported recurrent fusion genes in breast cancer that are the result of genomic rearrangements (13, 15, 16, 18, 34), read-through fusion transcripts in breast cancer have not been previously reported. We performed RNA-seq (35) on a total of 168 human samples, including 28 breast cancer cell lines, 42 fresh frozen triple negative breast cancer (TNBC) primary tumors, 42 fresh frozen estrogen receptor positive (ER+) breast cancer primary tumors, 21 fresh frozen non-neoplastic breast tissue samples that were adjacent to TNBC tumors, 30 fresh frozen non-neoplastic breast tissue samples that were adjacent to ER+ breast tumors, and five fresh frozen normal breast tissue samples that were collected from cancer-free patients during reduction mammoplasty procedures. We also downloaded RNA-seq data from 13 non-neoplastic human tissues collected by the Illumina Human Body Map 2.0 project, which includes adipose, brain, breast, colon, heart, kidney, liver, ovary, prostate, skeletal muscle, testes, thyroid and white blood cells (15). We used the ChimeraScan software package to identify read-through transcripts in the RNA-seq data (36).

We identified 17 candidate read-through fusion transcripts that were supported by at least 10 read-pairs that connect adjacent genes and at least one read that spanned the fusion junction in more than two breast cancer samples. We determined that six fusions were also detected in one or more of the non-neoplastic tissues from the Illumina Human Body Map 2.0 project and three transcripts were assigned to pairs of putative transcripts whose boundaries are not well defined. The remaining eight transcripts are breast tissue-specific read-through fusion transcripts. Read-through fusion transcripts with fusion junction-spanning reads are depicted in Figure 1, and the number of fusion junction-spanning reads in each sample is reported in Supplemental Table 1. For each of the eight fusion transcripts, we determined how many samples had at least one fusion junction-spanning read out of a collection of breast cancer cell lines, TNBC primary tumors, ER+ primary tumors, normal uninvolved tissue adjacent to each primary tumor type, and cancer-free normal tissue from reduction mammoplasty procedures (Table 1). To determine which read-through fusion transcripts were associated with breast cancer we used Fisher's Exact test to identify fusions that were significantly overrepresented in the breast cancer samples compared to the non-cancer breast samples (Table 1). Five of the read-through fusion transcripts were found at high frequency in normal breast tissue and were not significantly associated with breast cancer (*KLF16-REXO1, VAX2-ATP6V1B1, LOC100132832-CCDC146, MFGE8-HAPLN3, and CACNG4-CACNG1*; Table 1).

Three read-through fusion transcripts were significantly associated with breast cancer (*IL17RC-CRELD1*, *SCNN1A-TNFRSF1A and CTSD-IFITM10;* Fisher's Exact Test p-values in Table 1). Two of these breast cancer associated fusion transcripts were detected across breast tumors but were also detected at a lower frequency in normal uninvolved tissue that was adjacent to the primary tumors (*IL17RC-CRELD1*, and *SCNN1A-TNFRSF1A*) (Table 1). The breast tumors underwent macro-dissection to enrich for tumor cells; however, the adjacent normal uninvolved tissue was not dissected. Pathologists used a quality control section to diagnose the uninvolved tissue, but the specimen could have had infiltrating tumor cells or tumor exosomes containing mRNA deeper within the specimen. Neither of these fusions was detected in the cancer-free normal breast tissues from reduction mammoplasty procedures, suggesting that the low frequency of these fusions in the normal uninvolved tissue adjacent to tumors could be due to field defects. One fusion transcript, *CTSD-IFITM10,* was identified exclusively in breast cancer samples. All three of the breast cancer associated fusions were present in both ER+ and TNBC, and while they are present in different frequencies between the breast cancer subtypes, none are exclusive to a particular subtype. The breast cancer associated read-through fusion transcript are frequent events; 50% (14/28) of the breast cancer cell lines, 43% (18/42) of the TNBC primary tumors, and 24% (10/42) of the ER+ breast cancer primary tumors contained at least one of the three fusions.

All three of the breast cancer associated read-through fusion transcripts are spliced together using the last splice donor from the 5′ gene partner and the first splice acceptor in the 3′ gene partner, skipping the last exon of the 5′ gene

partner and the first exon of the 3′ gene partner (Figure 1). In each case, this results in splicing together nearly the full-length transcripts for both genes. Normally the 5′ fusion partner's transcript should be terminated by cleavage of the nascent transcript followed by polyadenylation (37). These read-through fusion transcripts have not been cleaved at the 5′ partner gene's polyadenylation signal and the 5′ partner gene's terminal exon splice acceptor site has been skipped to allow splicing between the adjacent genes. It is increasingly evident that the processes of transcription, splicing, 3′ transcript cleavage, and polyadenylation are coupled (38). One possible explanation for the generation of read-through fusion transcripts is that the 5′ partner gene's terminal exon was skipped because of a mutation at the splice acceptor site, which could hinder formation of the 3′-terminal exon-definition complex and subsequent cleavage/polyadenylation. If this were to occur, then the next available splice acceptor site would be at the 3′ partner gene's 2nd exon, consistent with the observed splice junctions. To test this notion, we PCR-amplified 200 bp surrounding the 5′ fusion partner gene's skipped splice acceptor site from DNA of cell lines with and without the fusion transcripts and sequenced the amplicons on an Illumina MiSeq machine. We did not identify any mutations at or near the splice acceptor sites associated with the presence of the fusion transcripts. We did observe both alleles of heterozygous SNPs at expected frequencies, so deletion of the splice sites is also unlikely. Because the processes of transcription, splicing, 3′ transcript cleavage, and polyadenylation can act both synergistically and competitively (38), it is possible that the kinetics of transcription at these loci is disrupted in breast cancer cells in a way that allows the formation of read-through fusion transcripts.

To determine if the expression level of each fusion partner gene was correlated with the presence of the fusion, we quantified the sequencing read depth for the canonical transcripts and fusion transcripts. In each of the samples containing fusion junction-spanning reads, we calculated the fraction of reads near the fusion junction that include sequence from the fusion transcript rather than the un-fused canonical transcripts (Figure 2a). In each case, the fraction of reads from the 3′ fusion partner that is involved in the fusion is significantly higher than the fraction of reads from the 5′ fusion partner that is participating in the fusion (Mann Whitney test: *IL17RC* vs *CRELD1* $p<0.0001$, *SCNN1A* vs *TNFRSF1A* $p = 0.0247$, and *CTSD* vs *IFITM10* $p <0.0001$). This indicates that a larger proportion of the transcription of the 3′ partner is created from read-through transcripts, and the promoter of the 5′ fusion partner likely regulates this expression. We examined the expression of the 5′ fusion partner in samples with and without evidence of the fusion transcript. We found that there was no difference in expression levels of *IL17RC, SCNN1A* or *TNFRSF1A* between samples with and without the fusion, indicating that the expression level of the 5′ gene partner is not associated with the presence of these fusions nor was the expression level of the 5′ gene partner associated with our power to detect the fusion (Figure 2b). The presence of the fusion transcripts is independent of the expression level of

the partner genes suggesting that other factors are responsible for their creation and regulation.

All three of the breast cancer associated read-through fusion transcripts we identified involved genes that encode membrane proteins. These proteins' functions rely on their correct placement in the membrane and correct participation in protein complexes. IL17RC is a single-pass type I membrane protein that binds the proinflammatory cytokines, IL-17A and IL-17F (39). It is fused to CRELD1, a membrane protein that contains an epidermal growth factor-like domain and is thought to function as a cell adhesion molecule (40). SCNN1A is an alpha subunit of nonvoltage-gated, amiloride-sensitive, sodium channels (41). It is fused to TNFRSF1A, a tumor necrosis factor-alpha receptor that activates NF-kappaB, mediates apoptosis, and regulates inflammatory responses (42). CTSD is a lysosomal aspartyl protease that also functions as a secreted protein that binds membrane receptors and has previously been associated with breast cancer (43). It is fused to IFITM10, a member of a family of membrane proteins that are induced by interferon and are involved in cell proliferation and cell adhesion (44). All of these read-through fusion transcripts join genes that have disparate functions, suggesting that a fused protein could impair normal function in breast cancer.

We predicted the length of the fusion protein based upon the location of the inter-gene splicing, and used Western blots with an antibody raised against one of the native partner proteins to determine whether a protein of the predicted fusion size could be detected in cell lysates from cell lines with and without RNA transcript evidence of the fusion. We observed specific Western blots of the targeted protein at the expected canonical size and detected protein at the predicted fusion size specifically in the cell lines with the fusion transcripts, and not in cell lines without the fusions for all three of the breast cancer associated read-through fusion transcripts (Figure 3). The cell line with the most fusion-spanning reads was positive for the fusion in all three Western blots, and in the case of the SCNN1A-TNFRSF1A, the cell line with the second highest number of fusion-spanning reads, was also positive by Western blot. These results suggest that the breast cancer associated read-through fusion transcripts are translated into fusion proteins.

To our knowledge, this is the first report of recurrent read-through fusion transcripts associated with breast cancer. Significant effort has been devoted to identifying gene expression and DNA mutations in breast cancer, and this reports adds aberrant mRNA read-through fusions to the list of molecular defects associated with the disease. Three recurrent fusion transcripts were associated with breast cancer, and for each of these, Western blots provided evidence of fusion proteins. The breast cancer associated read-through fusions involved membrane proteins, and represent exciting candidate biomarkers and potential therapeutic targets for further investigation. Future work to elucidate the mechanisms leading to the read-through transcription, mis-splicing, and loss of polyadenylation that create these fusions is also warranted to determine whether

a defect in the regulation of these processes is responsible for these aberrant transcripts.

Materials and methods:

Cell lines and tissues:

The 28 breast cancer cell lines were cultured as described previously (45). De-identified fresh frozen breast cancer specimens, fresh frozen breast tissue adjacent to tumors, and fresh frozen breast tissue specimens from reduction mammoplasty procedures were obtained from the University of Alabama at Birmingham's Comprehensive Cancer Center Tissue Procurement Shared Facility. The specific aliquots of specimens provided for research were chosen based on their quality control by board certified pathologists. After identification by quality control, the normal uninvolved breast tissue aliquots were not further macro-dissected. The breast tumor specimens were macro-dissected by the pathologists at the Tissue Procurement Shared Facility to enrich for tumor cell content and remove adjacent normal tissue. The frozen breast tissue specimens were weighed, transferred to a 15 mL conical tube containing ceramic beads, and RLT Buffer (Qiagen) plus 1% BME was added so that the tube contained 35 uL of buffer for each milligram of tissue. The conical tubes containing tissue, ceramic beads and buffer were then shaken in a MP Biomedicals FastPrep machine until the tissue was visibly homogenized (90 seconds at 6.5 meters per second). The homogenized tissue was stored at -80°C.

RNA-seq:

Total RNA was extracted from 5 million cultured cells or 350 uL of tissue homogenate (equivalent to 10 mg of tissue) using the Norgen Animal Tissue RNA Purification Kit (Norgen Biotek Corporation). Cell lysate was treated with Proteinase K before it was applied to the column and on-column DNAse treatment was performed according to the manufacturer's instructions. Total RNA was eluted from the columns and quantified using the Qubit RNA Assay Kit and the Qubit 2.0 fluorometer (Invitrogen). RNA-seq libraries for each sample were constructed from 250 ng total RNA using the polyA selection and transposase-based non-stranded library construction (Tn-RNA-seq) described previously (35). RNA-seq libraries were barcoded during PCR using Nextera barcoded primers according to the manufacturer (Epicentre). The RNA-seq libraries were quantified using the Qubit dsDNA HS Assay Kit and the Qubit 2.0 fluorometer (Invitrogen) and three barcoded libraries were pooled in equimolar quantities for sequencing. The pooled libraries were sequenced on an Illumina HiSeq 2000 sequencing machine using paired-end 50 bp reads and a 6 bp index read, and we obtained at least 50 million read pairs from each library. ChimeraScan 0.4.5a was used to align and identify fusion transcripts in each of the sequencing libraries using default parameters (36). To quantify the expression of each fusion partner, we used TopHat v1.4.1 (46) with the options –r 100 --mate-std-dev 75 to align 50 million RNA-seq read pairs, and used GENCODE version 9 (47) as a transcript reference. Gene expression values

(Fragments Per Kilobase of transcript Per Million reads, FPKMs) were calculated for each GENCODE transcript using Cufflinks 1.3.0 with the –u option (48).

Splice junction DNA sequencing:

Genomic DNA was isolated from 12 breast cancer cell lines using 5 million cultured cells per cell line and the Qiagen DNeasy Kit. PCR amplification of 200 bp surrounding the terminal exon splice acceptor site that is skipped in the formation of the read-through fusion transcripts were performed in 50 uL reactions containing 5 ng genomic DNA, 0.5 uM Forward PCR primer, 0.5 uM Reverse PCR primer, 5 units Platinum Taq DNA Polymerase (Invitrogen), 1x PCR Buffer with 2 mM $MgCl_2$, 0.5 mM each dNTP, and 0.5 M Betaine. These reactions were denatured at 98°C for 1 minute then thermocycled (30 cycles of 95°C for 30 seconds and 62°C for 3 minutes) and held at 4°C. The PCR products were purified using Agencourt AMPure XP beads (Beckman Coulter). The PCR products were quantified using the Qubit dsDNA HS Assay Kit and the Qubit 2.0 fluorometer (Invitrogen). Equimolar quantities of each of the eight PCR products were pooled into 12 pools, one for each cell line. Illumina sequencing libraries were prepared for each of the 12 pools of PCR products using Nextera according to the manufacturer's instructions (Epicentre). The 12 libraries were quantified using the Qubit dsDNA HS Assay Kit and the Qubit 2.0 fluorometer (Invitrogen). Equimolar quantities of each library were pooled and diluted to 10 nM and sequenced using single-end 50 bp reads and a 6 base index read on the Illumina MiSeq sequencer. We obtained 6 million sequencing reads in total covering all 8 amplicons in each of the 12 breast cancer cell lines. Variants were identified by the GATK software on BaseSpace (Illumina) and BAM files were downloaded and inspected manually using IGV 2.0 (49).

Western blots:

Breast cancer cell pellets containing 2.5 million cells were lysed by adding 100 uL RIPA Buffer (1x PBS, 1% NP-40, 0.5% sodium deoxycholate, 0.1% SDS and Roche protease inhibitor cocktail) and passing the solution through a 21-gauge needle. The lysed cells were then centrifuged at 16,000 rcf for 15 minutes at 4°C, and the supernatant was collected and protein was quantified using the Qubit Protein Assay Kit and the Qubit 2.0 fluorometer (Invitrogen). Twenty micrograms of protein extract was loaded into a BioRad 12% SDS-polyacrylamide gel in 1x Tris/Glycine Buffer (BioRad). Magic Marker (Invitrogen) was used as a protein standard. The gel electrophoresis rig was partially immersed in an ice bath while it ran for 1.5 hours at 125 V. Proteins were transferred to a nitrocellulose membrane using the iBlot system (Invitrogen) for 7 minutes at 20 V. The membranes were washed (1x PBS with 0.05% Tween 20) and incubated in blocking buffer for 60 minutes (1x PBS with 0.05% Tween 20 and 5% w/v Instant Nonfat Dry Milk). The membranes were then incubated with primary antibody overnight at 4°C (1x PBS with 0.05% Tween 20, 1% w/v Instant Nonfat Dry Milk, and 500 ng/mL primary antibody) followed by three 10 minute washes (1x PBS with 0.05% Tween 20). The following primary antibodies from Santa Cruz Biotechnology were used: CRELD1 sc-99364, CTSD sc-37438, and

TNFRSF1A sc-8436. The membrane was then incubated with secondary antibody (1x PBS, 0.05% Tween 20, 1% Instant Nonfat Dry Milk, and a 1:4,000 dilution of horseradish peroxidase (HRP) conjugated goat anti-mouse secondary antibody (Thermo Scientific)). The membrane was then washed (1x PBS with 0.05% Tween 20) and incubated for 5 minutes in a substrate solution of equal parts stable peroxide and luminol/enhancer (SuperSignal West Femto Chemiluminescent Substrate, Thermo Scientific). The membranes were then imaged for chemiluminescence.

Acknowledgements:

This study was supported in part by funding from TATRC, USAMRMC (W81XWH1010790), a Komen for the Cure Promise Grant (KG090969), and the National Institutes of Health, National Cancer Institute Specialized Program of Research Excellence (SPORE) in Breast Cancer (P50CA089019).

Tables and Figure Legends:

Table 1. Read-through fusion transcripts detected in breast samples. For each fusion transcript the number of samples containing junction-spanning reads is listed. Read-through fusion transcripts that are significantly associated with breast cancer are shaded in pink and p-values are listed in the last column.

Fusion Transcripts	Breast Cancer Cell Lines (N=28)	TNBC Primary Tumors (N=42)	ER+ Breast Cancer Primary Tumors (N=42)	Normal Uninvolved Tissue Adjacent to TNBC (N=21)	Normal Uninvolved Tissue Adjacent to ER+ Breast Cancer (N=30)	Cancer-Free Reduction Mammoplasty Breast Tissue (N=5)	Human Body Map (N=13)	Cancer vs. Normal Fisher's Exact Test p-value
KLF16-REXO1	7 (25%)	18 (43%)	16 (38%)	14 (67%)	15 (50%)	2 (40%)	0 (0%)	0.1699*
VAX2-ATP6V1B1	6 (21%)	8 (19%)	4 (10%)	4 (19%)	3 (10%)	2 (40%)	0 (0%)	0.3711
LOC100132832-CCDC146	2 (7%)	11 (26%)	5 (12%)	5 (24%)	3 (10%)	1 (20%)	0 (0%)	0.3711
MFGE8-HAPLN3	4 (14%)	23 (55%)	4 (10%)	4 (19%)	7 (23%)	1 (20%)	0 (0%)	0.0794
CACNG4-CACNG1	2 (7%)	2 (5%)	12 (29%)	1 (5%)	3 (10%)	0 (0%)	0 (0%)	0.0600
IL17RC-CRELD1	3 (11%)	11 (26%)	4 (10%)	3 (14%)	1 (3%)	0 (0%)	0 (0%)	0.0306
SCNN1A-TNFRSF1A	10 (36%)	3 (7%)	5 (12%)	1 (5%)	1 (3%)	0 (0%)	0 (0%)	0.0039
CTSD-IFITM10	7 (25%)	9 (21%)	5 (12%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	< 0.0001

* More prevalent in non-cancer samples.

Figure Legends:

Figure 1. Read-through fusion transcripts identified in breast samples. Eight read-through fusion transcripts were detected in more than two breast samples using paired-end RNA-seq. These read-through fusions were breast-tissue specific, and not detected in other non-neoplastic human tissues sequenced by the Illumina Human Body Map 2.0 project. The exon structure of the 5′ gene partner is depicted in green, and the exon structure of the 3′ gene

partner is depicted in red. The fusion transcripts use endogenous splice sites and black lines indicate which exons flank the fusion junction to result in the chimeric transcript. RNA-seq reads that span the fusion junction are depicted above the gene models and the sequence from the 5′ partner is in green text and the sequence from the 3′ partner is in red text. The intergenic chromosomal distance between the fusion partners is denoted in kilobase pairs (kbp). The five read-through fusion transcripts depicted in a, b, c, d and e were detected in both breast cancer specimens and non-cancer breast tissue. Three read-through fusion transcripts significantly associated with breast cancer are depicted in f, g and h.

Figure 2. Expression of fusion partners for breast cancer associated read-through fusion transcripts. a) We computed the fraction of reads near the fusion junction that include sequence from the fusion transcript rather than the un-fused canonical transcript. The fraction of fusion transcript reads for 5′ fusion partners are indicated in green and the 3′ fusion partners are denoted in red. Mean and standard error of the mean are depicted in black. Less than 20% of the 5′ fusion partners' transcripts have the fusion sequence, indicating that most of the transcripts from the 5′ fusion partners are not fused. A significantly larger fraction of the 3′ fusion partners' transcripts contain the fusion sequence. This indicates that the expression of the 3′ fusion partner is composed of a large fraction of fusion transcript driven by the 5′ fusion partner's promoter. b) There is no difference in the expression levels (Fragments Per Kilobase of transcript Per Million reads; FPKMs) of the 5′ fusion partner between samples with or without the read-through fusion transcript (labeled Fused and Not Fused, respectively). Mean and standard error of the mean are depicted in black. This indicates that increased expression of the 5′ fusion partner is not sufficient to induce read-through fusion transcripts, and that lower expression of the 5′ partner is not associated with our power to detect the read-through fusion transcripts.

Figure 3. Western blots of three breast cancer associated fusion proteins. We performed Western blots using antibodies raised to one of the fusion partner proteins for the three breast cancer associated fusion transcripts. For each candidate fusion, we ran cell lysates from two cell lines with RNA-seq reads spanning the fusion junction and one cell line without RNA-seq reads spanning the fusion junction. In each blot, the canonical/native size of the targeted protein was detected in each cell line, and a band at the predicted fusion protein size was detected in the cell line with the most RNA-seq fusion-spanning reads (IL17RC-CRELD1 in SUM-149, CTSD-IFITM10 in MCF7, and SCNN1A-TNFRSF1A in HCC1954). A band corresponding to the size of the predicted fusion protein was also detected in the cell line with the second most RNA-seq fusion transcript reads for the SCNN1A-TNFRSF1A fusion (SUM-102). None of the cell lines without RNA-seq evidence of the fusion transcript produced fusion protein-sized bands.

References:

1. Nowell PC. The minute chromosome (Phl) in chronic granulocytic leukemia. Blut. 1962 Apr;8:65-6.

2. Rowley JD. Letter: A new consistent chromosomal abnormality in chronic myelogenous leukaemia identified by quinacrine fluorescence and Giemsa staining. Nature. 1973 Jun 1;243(5405):290-3.

3. Rowley JD. Chromosomal translocations: revisited yet again. Blood. 2008 Sep 15;112(6):2183-9.

4. Druker BJ, Tamura S, Buchdunger E, Ohno S, Segal GM, Fanning S, et al. Effects of a selective inhibitor of the Abl tyrosine kinase on the growth of Bcr-Abl positive cells. Nat Med. 1996 May;2(5):561-6.

5. Mitelman F, Johansson B, Mertens F. Fusion genes and rearranged genes as a linear function of chromosome aberrations in cancer. Nat Genet. 2004 Apr;36(4):331-4.

6. Maher CA, Palanisamy N, Brenner JC, Cao X, Kalyana-Sundaram S, Luo S, et al. Chimeric transcript discovery by paired-end transcriptome sequencing. Proc Natl Acad Sci U S A. 2009 Jul 28;106(30):12353-8.

7. Tognon C, Knezevich SR, Huntsman D, Roskelley CD, Melnyk N, Mathers JA, et al. Expression of the ETV6-NTRK3 gene fusion as a primary event in human secretory breast carcinoma. Cancer Cell. 2002 Nov;2(5):367-76.

8. Persson M, Andren Y, Mark J, Horlings HM, Persson F, Stenman G. Recurrent fusion of MYB and NFIB transcription factor genes in carcinomas of the breast and head and neck. Proc Natl Acad Sci U S A. 2009 Nov 3;106(44):18740-4.

9. Tomlins SA, Laxman B, Dhanasekaran SM, Helgeson BE, Cao X, Morris DS, et al. Distinct classes of chromosomal rearrangements create oncogenic ETS gene fusions in prostate cancer. Nature. 2007 Aug 2;448(7153):595-9.

10. Tomlins SA, Rhodes DR, Perner S, Dhanasekaran SM, Mehra R, Sun XW, et al. Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer. Science. 2005 Oct 28;310(5748):644-8.

11. Kumar-Sinha C, Tomlins SA, Chinnaiyan AM. Recurrent gene fusions in prostate cancer. Nat Rev Cancer. 2008 Jul;8(7):497-511.

12. Soda M, Choi YL, Enomoto M, Takada S, Yamashita Y, Ishikawa S, et al. Identification of the transforming EML4-ALK fusion gene in non-small-cell lung cancer. Nature. 2007 Aug 2;448(7153):561-6.

13. Robinson DR, Kalyana-Sundaram S, Wu YM, Shankar S, Cao X, Ateeq B, et al. Functionally recurrent rearrangements of the MAST kinase and Notch gene families in breast cancer. Nat Med. 2011 Dec;17(12):1646-51.

14. Asmann YW, Hossain A, Necela BM, Middha S, Kalari KR, Sun Z, et al. A novel bioinformatics pipeline for identification and characterization of fusion transcripts in breast cancer and normal cell lines. Nucleic Acids Res. 2011 Aug;39(15):e100.

15. Asmann YW, Necela BM, Kalari KR, Hossain A, Baker TR, Carr JM, et al. Detection of redundant fusion transcripts as biomarkers or disease-specific therapeutic targets in breast cancer. Cancer Res. 2012 Apr 15;72(8):1921-8.

16. Edgren H, Murumagi A, Kangaspeska S, Nicorici D, Hongisto V, Kleivi K, et al. Identification of fusion genes in breast cancer by paired-end RNA-sequencing. Genome Biol. 2011;12(1):R6.

17. Ha KC, Lalonde E, Li L, Cavallone L, Natrajan R, Lambros MB, et al. Identification of gene fusion transcripts by transcriptome sequencing in BRCA1-mutated breast cancers and cell lines. BMC Med Genomics. 2011;4:75.

18. Zhao Q, Caballero OL, Levy S, Stevenson BJ, Iseli C, de Souza SJ, et al. Transcriptome-guided characterization of genomic rearrangements in a breast cancer cell line. Proc Natl Acad Sci U S A. 2009 Feb 10;106(6):1886-91.

19. Kim D, Salzberg SL. TopHat-Fusion: an algorithm for discovery of novel fusion transcripts. Genome Biol. 2011;12(8):R72.

20. Frenkel-Morgenstern M, Lacroix V, Ezkurdia I, Levin Y, Gabashvili A, Prilusky J, et al. Chimeras taking shape: potential functions of proteins encoded by chimeric RNA transcripts. Genome Res. 2012 Jul;22(7):1231-42.

21. Li X, Zhao L, Jiang H, Wang W. Short homologous sequences are strongly associated with the generation of chimeric RNAs in eukaryotes. J Mol Evol. 2009 Jan;68(1):56-65.

22. Akiva P, Toporik A, Edelheit S, Peretz Y, Diber A, Shemesh R, et al. Transcription-mediated gene fusion in the human genome. Genome Res. 2006 Jan;16(1):30-6.

23. Parra G, Reymond A, Dabbouseh N, Dermitzakis ET, Castelo R, Thomson TM, et al. Tandem chimerism as a means to increase protein complexity in the human genome. Genome Res. 2006 Jan;16(1):37-44.

24. Li H, Wang J, Ma X, Sklar J. Gene fusions and RNA trans-splicing in normal and neoplastic human cells. Cell Cycle. 2009 Jan 15;8(2):218-22.

25. Rickman DS, Pflueger D, Moss B, VanDoren VE, Chen CX, de la Taille A, et al. SLC45A3-ELK4 is a novel and frequent erythroblast transformation-specific fusion transcript in prostate cancer. Cancer Res. 2009 Apr 1;69(7):2734-8.

26. Kim RN, Kim A, Choi SH, Kim DS, Nam SH, Kim DW, et al. Novel mechanism of conjoined gene formation in the human genome. Funct Integr Genomics. 2012 Mar;12(1):45-61.

27. Prakash T, Sharma VK, Adati N, Ozawa R, Kumar N, Nishida Y, et al. Expression of conjoined genes: another mechanism for gene regulation in eukaryotes. PLoS One. 2010;5(10):e13284.

28. Kumar-Sinha C, Kalyana-Sundaram S, Chinnaiyan AM. SLC45A3-ELK4 Chimera in Prostate Cancer: Spotlight on cis-Splicing. Cancer Discov. 2012 Jul;2(7):582-5.

29. Nacu S, Yuan W, Kan Z, Bhatt D, Rivers CS, Stinson J, et al. Deep RNA sequencing analysis of readthrough gene fusions in human prostate adenocarcinoma and reference samples. BMC Med Genomics. 2011;4:11.

30. Maher CA, Kumar-Sinha C, Cao X, Kalyana-Sundaram S, Han B, Jing X, et al. Transcriptome sequencing to detect gene fusions in cancer. Nature. 2009 Mar 5;458(7234):97-101.

31. Zhang Y, Gong M, Yuan H, Park HG, Frierson HF, Li H. Chimeric Transcript Generated by cis-Splicing of Adjacent Genes Regulates Prostate Cancer Cell Proliferation. Cancer Discov. 2012 Jul;2(7):598-607.

32. Kannan K, Wang L, Wang J, Ittmann MM, Li W, Yen L. Recurrent chimeric RNAs enriched in human prostate cancer identified by deep sequencing. Proc Natl Acad Sci U S A. 2011 May 31;108(22):9172-7.

33. Zhou J, Liao J, Zheng X, Shen H. Chimeric RNAs as potential biomarkers for tumor diagnosis. BMB Rep. 2012 Mar;45(3):133-40.

34. Stephens PJ, McBride DJ, Lin ML, Varela I, Pleasance ED, Simpson JT, et al. Complex landscapes of somatic rearrangement in human breast cancer genomes. Nature. 2009 Dec 24;462(7276):1005-10.

35. Gertz J, Varley KE, Davis NS, Baas BJ, Goryshin IY, Vaidyanathan R, et al. Transposase mediated construction of RNA-seq libraries. Genome Res. 2012 Jan;22(1):134-41.

36. Iyer MK, Chinnaiyan AM, Maher CA. ChimeraScan: a tool for identifying chimeric transcription in sequencing data. Bioinformatics. 2011 Oct 15;27(20):2903-4.

37. Wahle E, Ruegsegger U. 3'-End processing of pre-mRNA in eukaryotes. FEMS Microbiol Rev. 1999 Jun;23(3):277-95.

38. Martinson HG. An active role for splicing in 3'-end formation. Wiley Interdiscip Rev RNA. 2011 Jul-Aug;2(4):459-70.

39. Kuestner RE, Taft DW, Haran A, Brandt CS, Brender T, Lum K, et al. Identification of the IL-17 receptor related molecule IL-17RC as the receptor for IL-17F. J Immunol. 2007 Oct 15;179(8):5462-73.

40. Rupp PA, Fouad GT, Egelston CA, Reifsteck CA, Olson SB, Knosp WM, et al. Identification, genomic organization and mRNA expression of CRELD1, the founding member of a unique family of matricellular proteins. Gene. 2002 Jun 26;293(1-2):47-57.

41. Hummler E, Beermann F. Scnn1 sodium channel gene family in genetically engineered mice. J Am Soc Nephrol. 2000 Nov;11 Suppl 16:S129-34.

42. Chen G, Goeddel DV. TNF-R1 signaling: a beautiful pathway. Science. 2002 May 31;296(5573):1634-5.

43. Nicotra G, Castino R, Follo C, Peracchio C, Valente G, Isidoro C. The dilemma: does tissue expression of cathepsin D reflect tumor malignancy? The question: does the assay truly mirror cathepsin D mis-function in the tumor? Cancer Biomark. 2010;7(1):47-64.

44. Hickford D, Frankenberg S, Shaw G, Renfree MB. Evolution of vertebrate interferon inducible transmembrane proteins. BMC Genomics. 2012;13:155.

45. Oliver PG, LoBuglio AF, Zhou T, Forero A, Kim H, Zinn KR, et al. Effect of anti-DR5 and chemotherapy on basal-like breast cancer. Breast Cancer Res Treat. 2012 Jun;133(2):417-26.

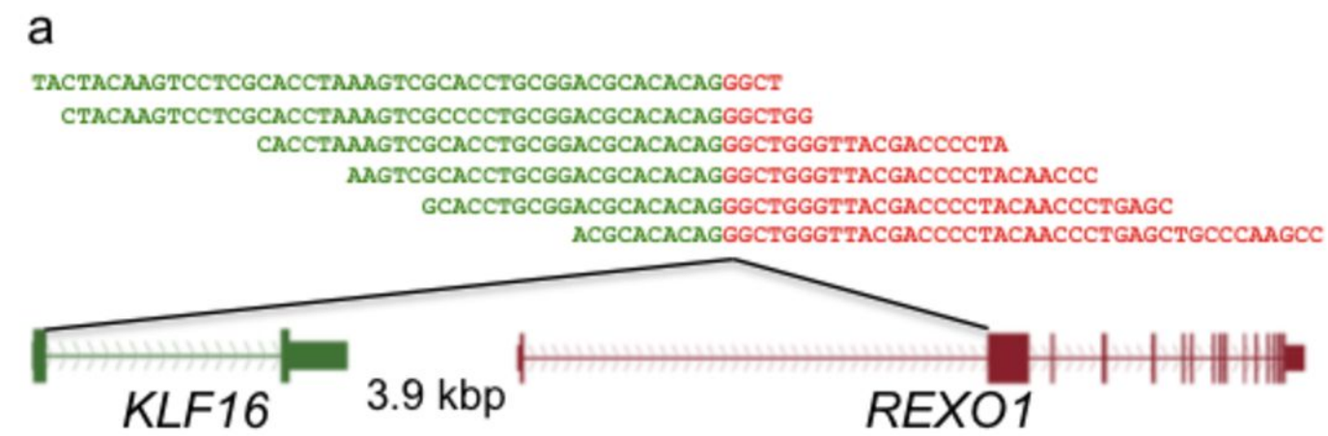
46. Trapnell C, Pachter L, Salzberg SL. TopHat: discovering splice junctions with RNA-Seq. Bioinformatics. 2009 May 1;25(9):1105-11.

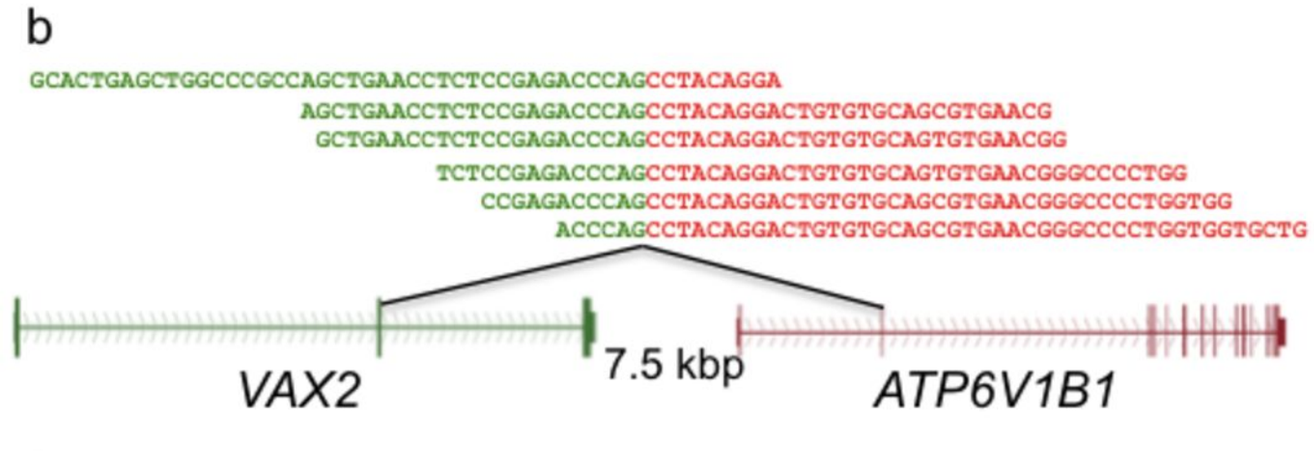
47. Harrow J, Denoeud F, Frankish A, Reymond A, Chen CK, Chrast J, et al. GENCODE: producing a reference annotation for ENCODE. Genome Biol. 2006;7 Suppl 1:S4 1-9.

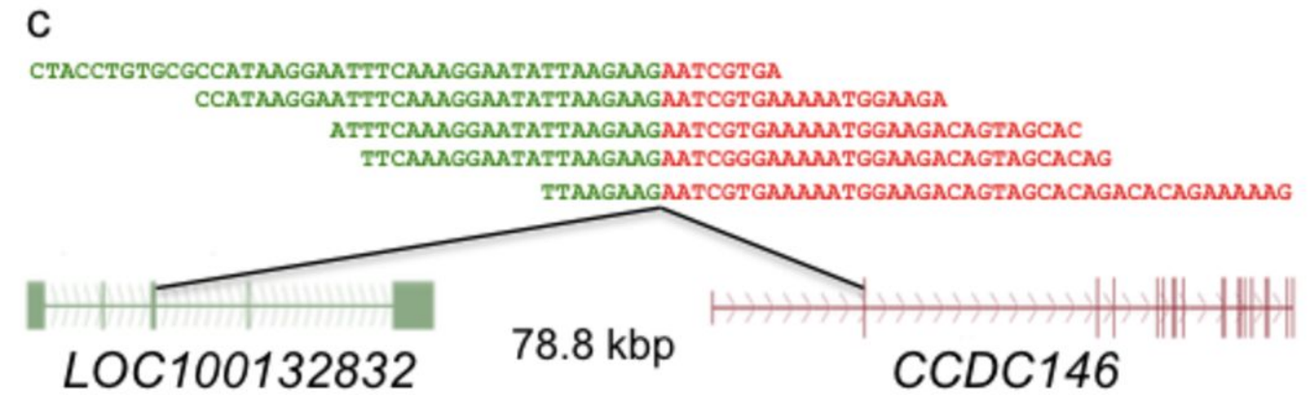
48. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol. 2010 May;28(5):511-5.

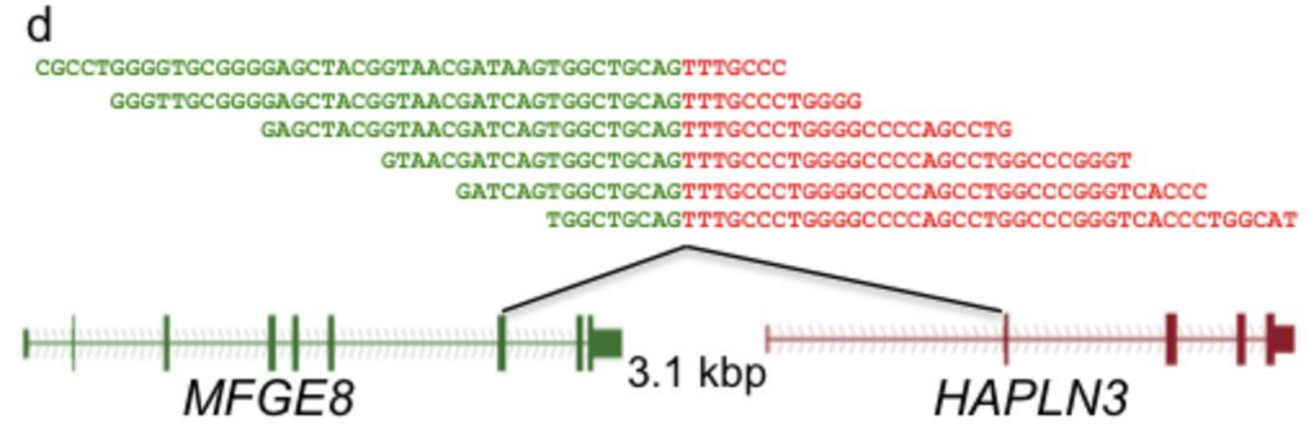
49. Robinson JT, Thorvaldsdottir H, Winckler W, Guttman M, Lander ES, Getz G, et al. Integrative genomics viewer. Nat Biotechnol. 2011 Jan;29(1):24-6.

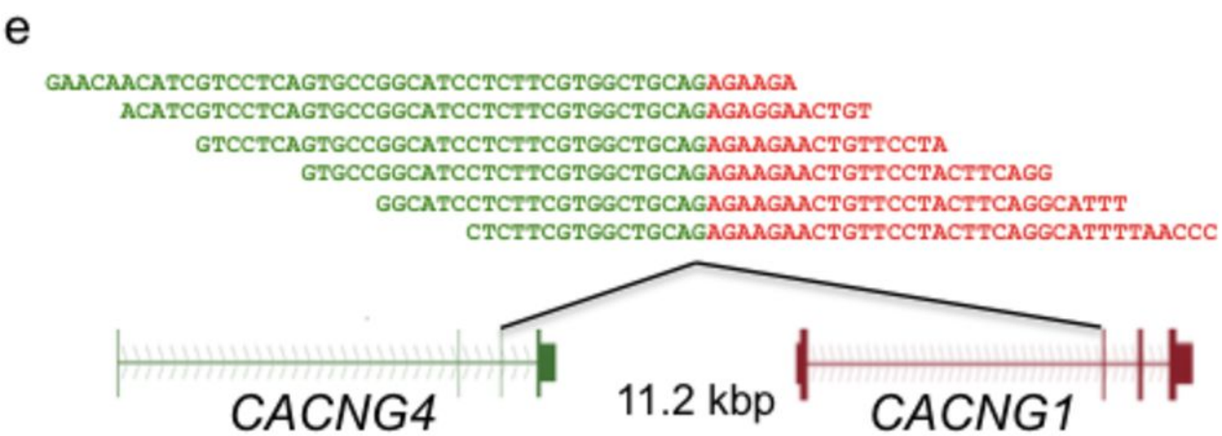
Figure 1

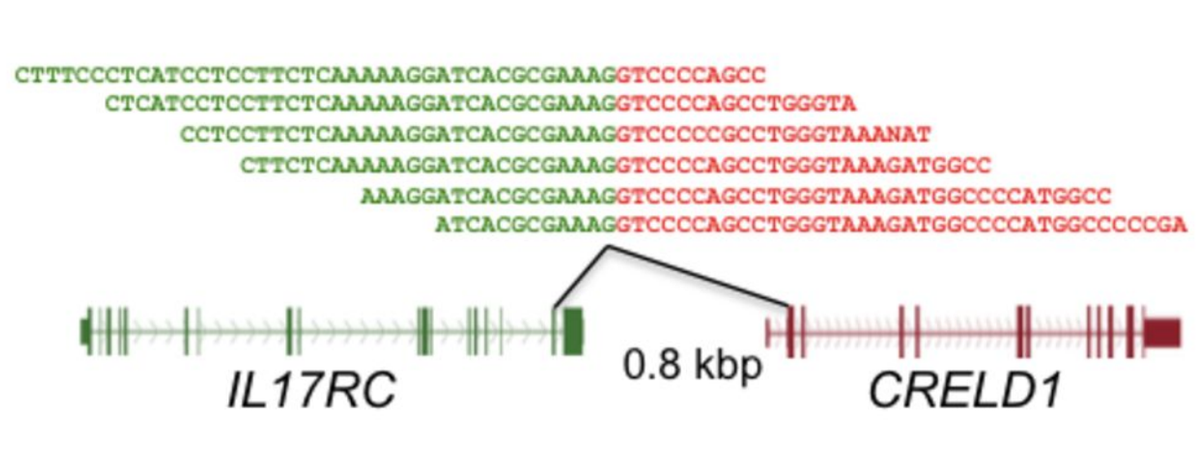

a
TACTACAAGTCCTCGCACCTAAAGTCGCACCTGCGGACGCACACAGGGCT
CTACAAGTCCTCGCACCTAAAGTCGCCCCTGCGGACGCACACAGGGCTGG
CACCTAAAGTCGCACCTGCGGACGCACACAGGGCTGGGTTACGACCCCTA
AAGTCGCACCTGCGGACGCACACAGGGCTGGGTTACGACCCCTACAACCC
GCACCTGCGGACGCACACAGGGCTGGGTTACGACCCCTACAACCCTGAGC
ACGCACACAGGGCTGGGTTACGACCCCTACAACCCTGAGCTGCCCAAGCC
KLF16
3.9 kbp
REXO1

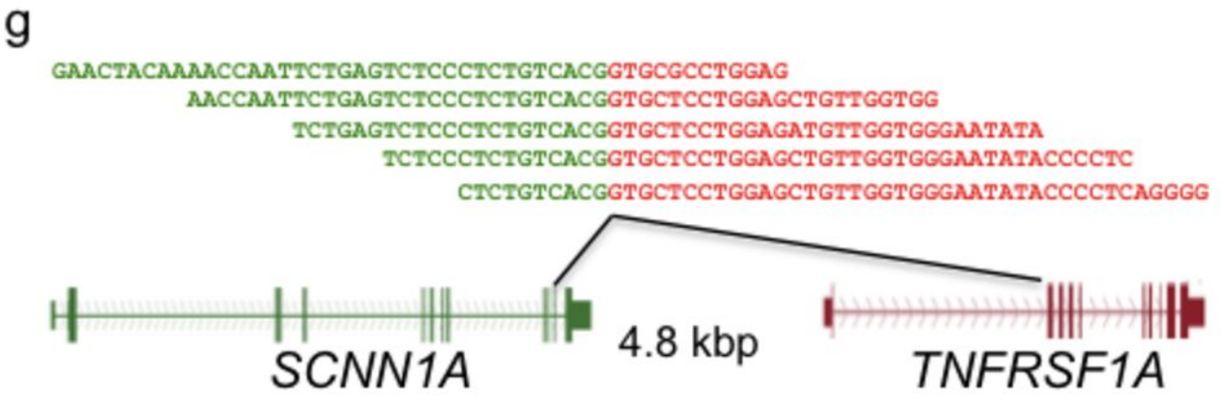

b
GCACTGAGCTGGCCCGCCAGCTGAACCTCTCCGAGACCCAGCCTACAGGA
AGCTGAACCTCTCCGAGACCCAGCCTACAGGACTGTGTGCAGCGTGAACG
GCTGAACCTCTCCGAGACCCAGCCTACAGGACTGTGTGCAGTGTGAACGG
TCTCCGAGACCCAGCCTACAGGACTGTGTGCAGTGTGAACGGGCCCCTGG
CCGAGACCCAGCCTACAGGACTGTGTGCAGCGTGAACGGGCCCCTGGTGG
ACCCAGCCTACAGGACTGTGTGCAGCGTGAACGGGCCCCTGGTGGTGCTG
VAX2
7.5 kbp
ATP6V1B1


c
CTACCTGTGCGCCATAAGGAATTTCAAAGGAATATTAAGAAGAATCGTGA
CCATAAGGAATTTCAAAGGAATATTAAGAAGAATCGTGAAAAAATGGAAGA
ATTTCAAAGGAATATTAAGAAGAATCGTGAAAAAATGGAAGACAGTAGCAC
TTCAAAGGAATATTAAGAAGAATCGGGAAAAAATGGAAGACAGTAGCACAG
TTAAGAAGAATCGTGAAAAAATGGAAGACAGTAGCACAGACACAGAAAAAG
LOC100132832
78.8 kbp
CCDC146


d
CGCCTGGGGTGCGGGGAGCTACGGTAACGATAAGTGGCTGCAGTTTGCCC
GGGTTGCGGGGAGCTACGGTAACGATCAGTGGCTGCAGTTTGCCCTGGGG
GAGCTACGGTAACGATCAGTGGCTGCAGTTTGCCCTGGGGCCCCAGCCTG
GTAACGATCAGTGGCTGCAGTTTGCCCTGGGGCCCCAGCCTGGCCCGGGT
GATCAGTGGCTGCAGTTTGCCCTGGGGCCCCAGCCTGGCCCGGGTCACCC
TGGCTGCAGTTTGCCCTGGGGCCCCAGCCTGGCCCGGGTCACCCTGGCAT
MFGE8
3.1 kbp
HAPLN3


e
GAACAACATCGTCCTCAGTGCCGGCATCCTCTTCGTGGCTGCAGAGAAGA
ACATCGTCCTCAGTGCCGGCATCCTCTTCGTGGCTGCAGAGAGGAACTGT
GTCCTCAGTGCCGGCATCCTCTTCGTGGCTGCAGAGAAGAACTGTTCCTA
GTGCCGGCATCCTCTTCGTGGCTGCAGAGAAGAACTGTTCCTACTTCAGG
GGCATCCTCTTCGTGGCTGCAGAGAAGAACTGTTCCTACTTCAGGCATTT
CTCTTCGTGGCTGCAGAGAAGAACTGTTCCTACTTCAGGCATTTTAACCC
CACNG4
11.2 kbp
CACNG1


f
CTTTCCCTCATCCTCCTTCTCAAAAAGGATCACGCGAAAGGTCCCCAGCC
CTCATCCTCCTTCTCAAAAAGGATCACGCGAAAGGTCCCCAGCCTGGGTA
CCTCCTTCTCAAAAAGGATCACGCGAAAGGTCCCCCGCCTGGGTAAANAT
CTTCTCAAAAAGGATCACGCGAAAGGTCCCCAGCCTGGGTAAAGATGGCC
AAAGGATCACGCGAAAGGTCCCCAGCCTGGGTAAAGATGGCCCCATGGCC
ATCACGCGAAAGGTCCCCAGCCTGGGTAAAGATGGCCCCATGGCCCCCGA
IL17RC
0.8 kbp
CRELD1


g
GAACTACAAAACCAATTCTGAGTCTCCCTCTGTCACGGTGCGCCTGGAG
AACCAATTCTGAGTCTCCCTCTGTCACGGTGCTCCTGGAGCTGTTGGTGG
TCTGAGTCTCCCTCTGTCACGGTGCTCCTGGAGATGTTGGTGGGAATATA
TCTCCCTCTGTCACGGTGCTCCTGGAGCTGTTGGTGGGAATATACCCCTC
CTCTGTCACGGTGCTCCTGGAGCTGTTGGTGGGAATATACCCCTCAGGGG
SCNN1A
4.8 kbp
TNFRSF1A

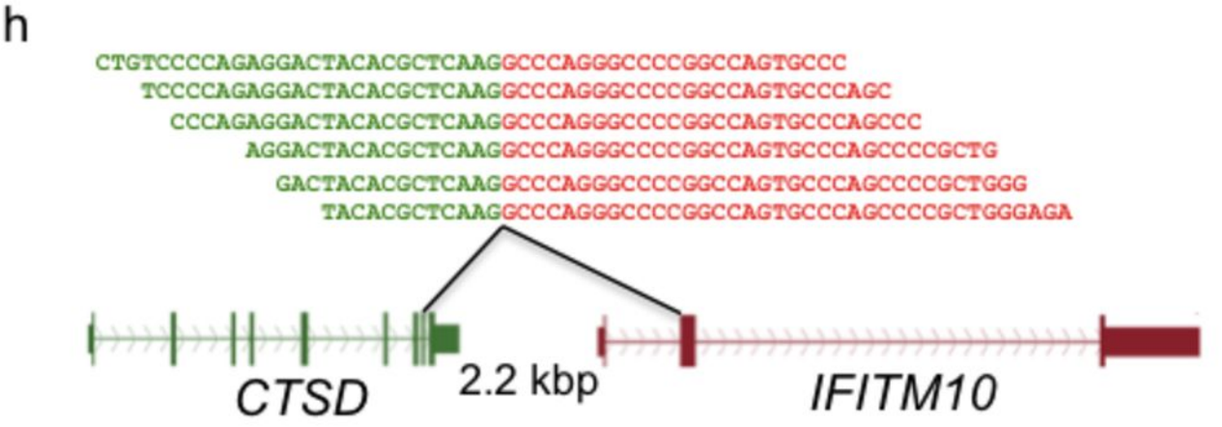
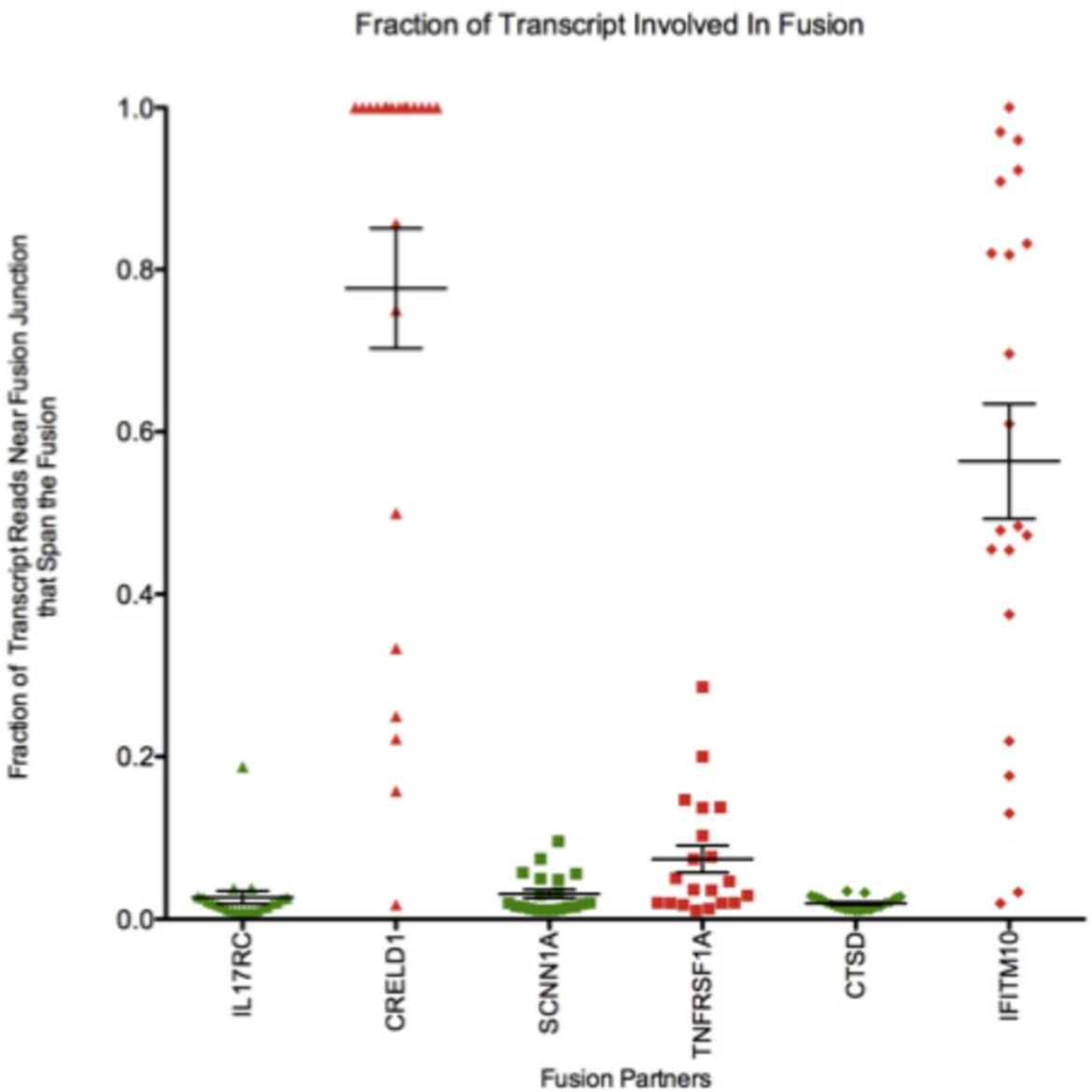

h
CTGTCCCCAGAGGACTACACGCTCAAGGCCCAGGGCCCCGGCCAGTGCCC
TCCCCAGAGGACTACACGCTCAAGGCCCAGGGCCCCGGCCAGTGCCCAGC
CCCAGAGGACTACACGCTCAAGGCCCAGGGCCCCGGCCAGTGCCCAGCCC
AGGACTACACGCTCAAGGCCCAGGGCCCCGGCCAGTGCCCAGCCCCGCTG
GACTACACGCTCAAGGCCCAGGGCCCCGGCCAGTGCCCAGCCCCGCTGGG
TACACGCTCAAGGCCCAGGGCCCCGGCCAGTGCCCAGCCCCGCTGGGAGA
CTSD
2.2 kbp
IFITM10

Figure 2

a


Fraction of Transcript Involved In Fusion
Fraction of Transcript Reads Near Fusion Junction that Span the Fusion
1.0
0.8
0.6
0.4
0.2
0.0
IL17RC
CRELD1
SCNN1A
TNFRSF1A
CTSD
IFITM10
Fusion Partners

b

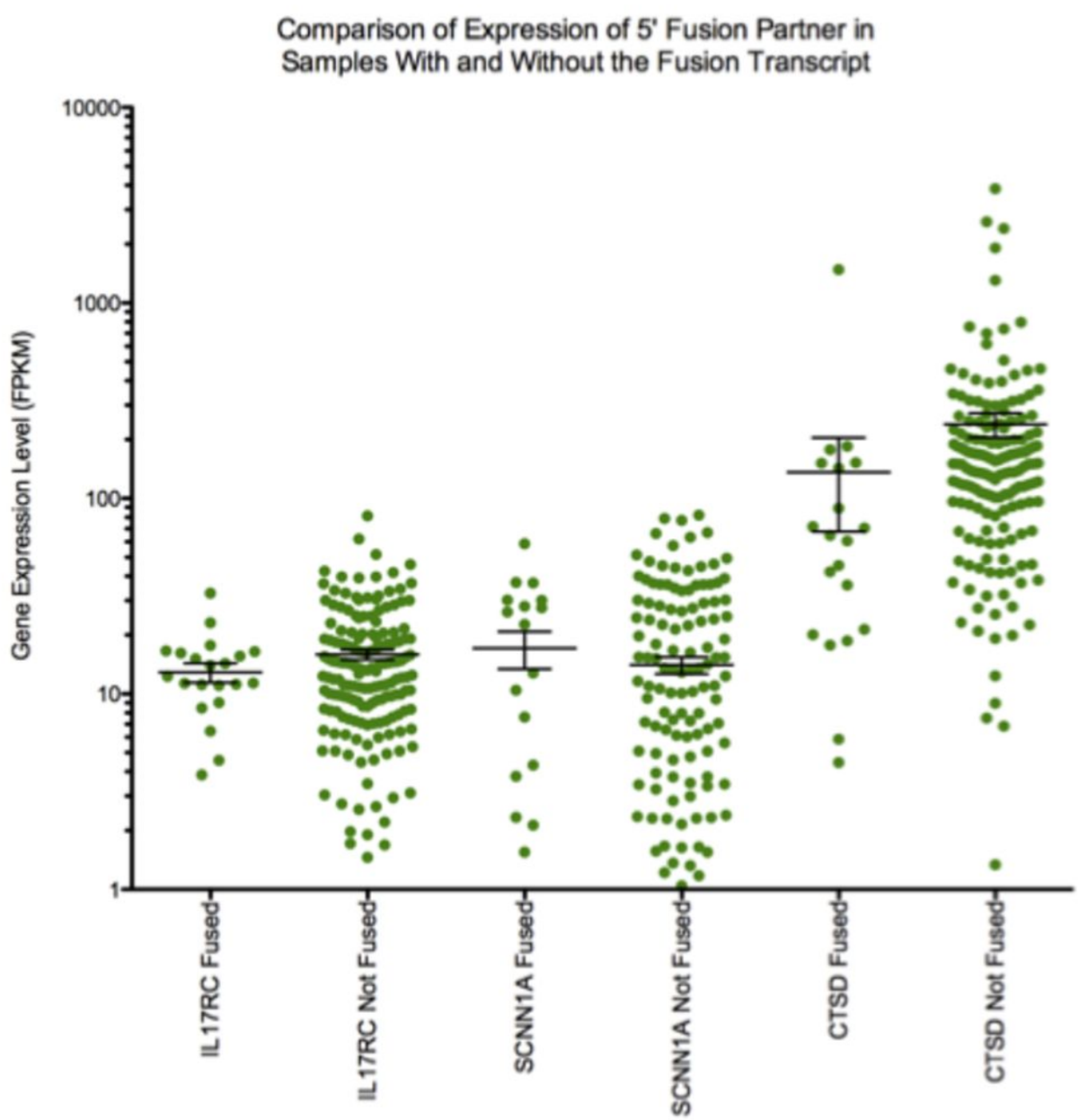
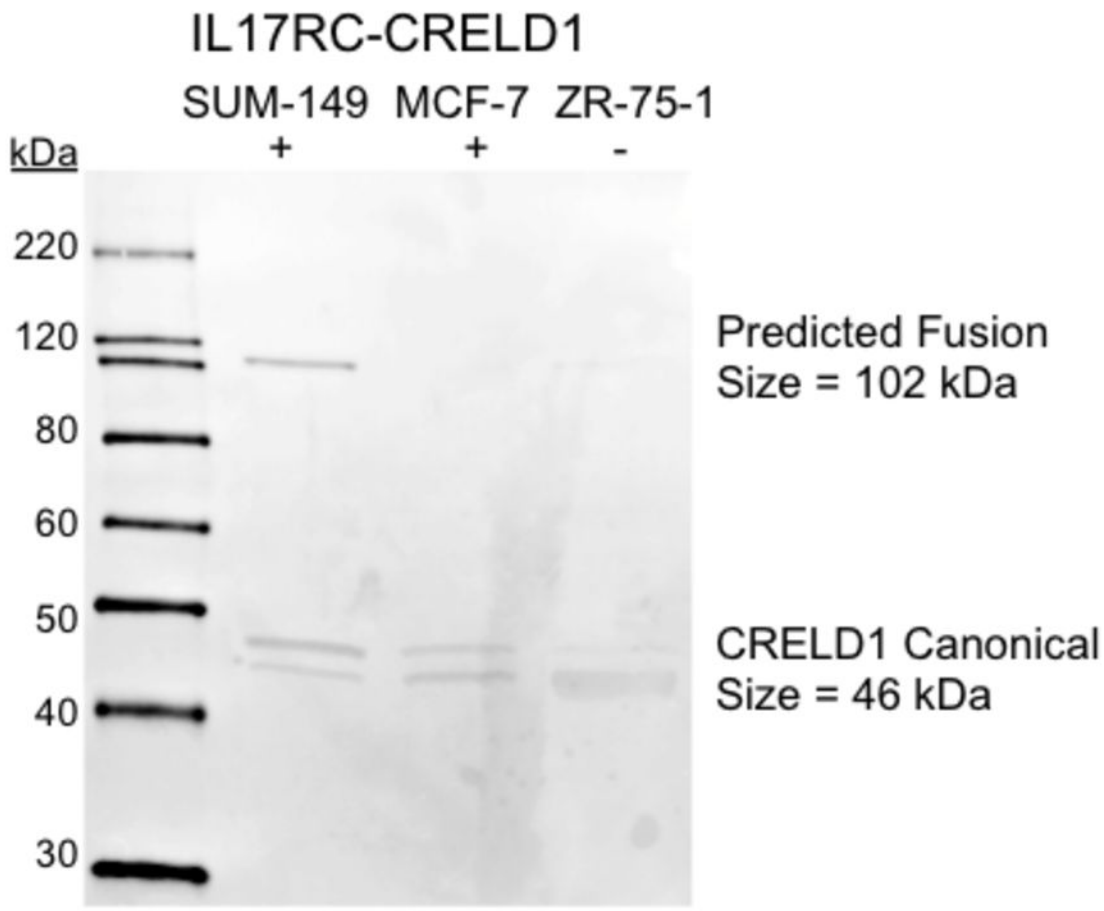
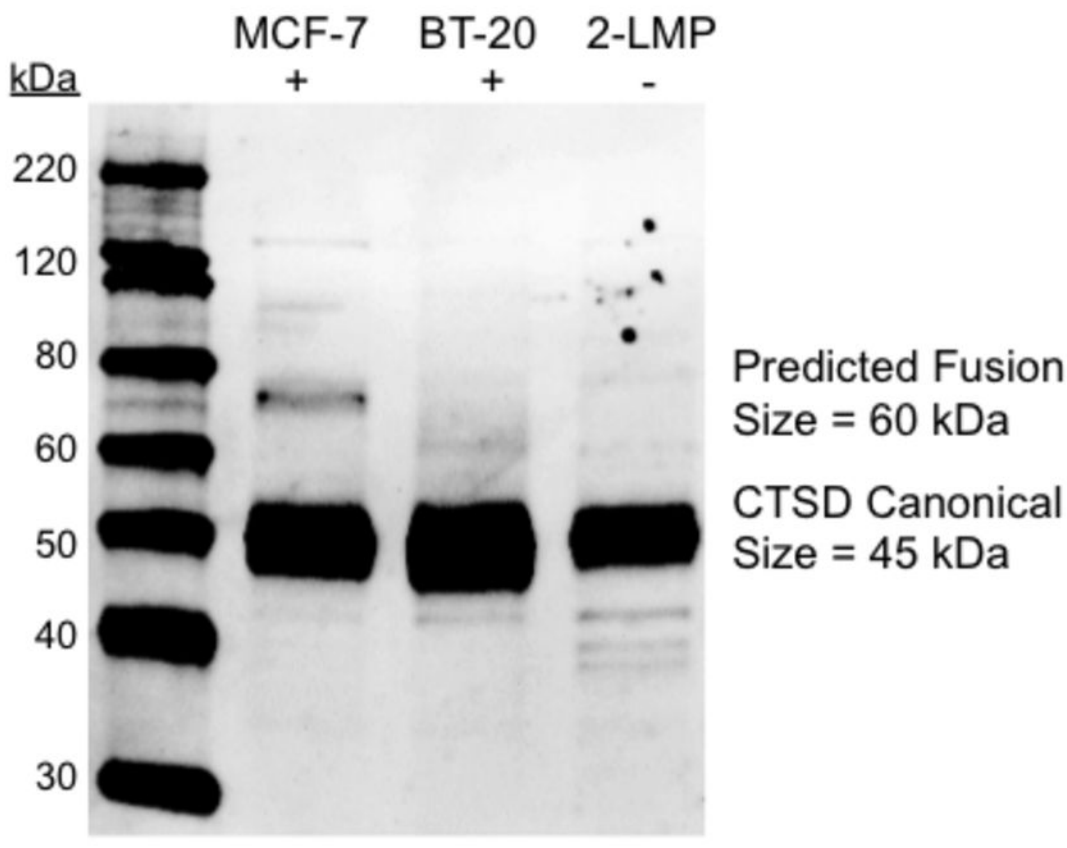
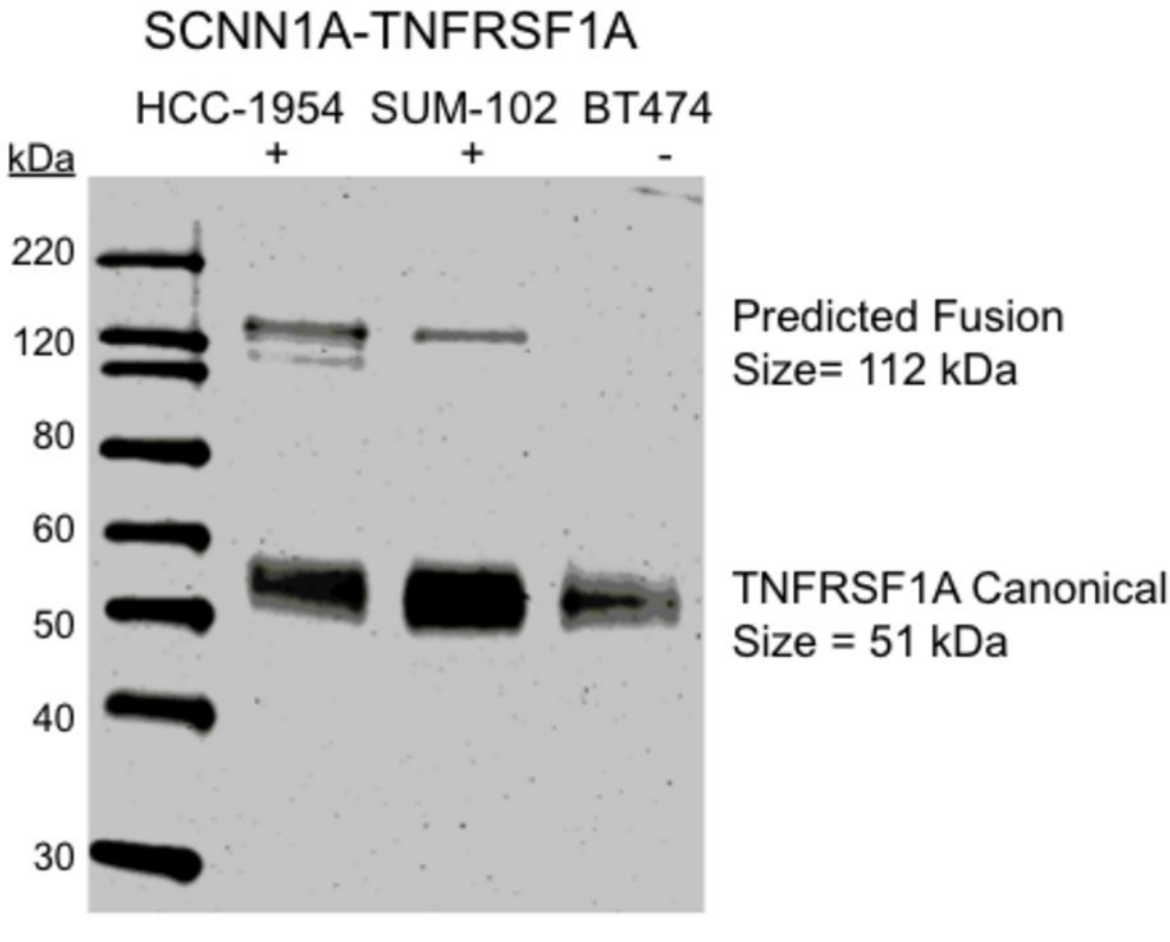

Comparison of Expression of 5' Fusion Partner in Samples With and Without the Fusion Transcript
Gene Expression Level (FPKM)
10000
1000
100
10
1
IL17RC Fused
IL17RC Not Fused
SCNN1A Fused
SCNN1A Not Fused
CTSD Fused
CTSD Not Fused

Figure 3


IL17RC-CRELD1
SUM-149 MCF-7 ZR-75-1
kDa
+ + -
220
120
80
60
50
40
30
Predicted Fusion
Size = 102 kDa
CRELD1 Canonical
Size = 46 kDa


CTSD-IFITM10
MCF-7 BT-20 2-LMP
kDa
+ + -
220
120
80
60
50
40
30
Predicted Fusion
Size = 60 kDa
CTSD Canonical
Size = 45 kDa


SCNN1A-TNFRSF1A
HCC-1954 SUM-102 BT474
kDa
+ + -
220
120
80
60
50
40
30
Predicted Fusion
Size= 112 kDa
TNFRSF1A Canonical
Size = 51 kDa

Appendix C

Genome-wide DNA methylation and RNA-seq analysis of tumor and normal prostate tissue for biomarker discovery

Marie K. Cross[1], Katherine E. Varley[1], Jason Gertz[1], Nick S. Davis[1], Devin M. Absher[1], James D. Brooks[2], and Richard M. Myers[1]

[1]HudsonAlpha Institute for Biotechnology, Huntsville, AL
[2]Stanford University, Palo Alto, CA

Prostate tumors frequently exhibit altered DNA methylation and gene expression patterns when compared to adjacent normal tissue. These disease-specific changes are promising candidate biomarkers due to their specificity and the sensitive detection of methylated DNA in peripheral fluids. There is a clear need to identify novel, noninvasive biomarkers for the diagnosis and prognosis of prostate cancer. Current diagnostic tools for prostate cancer lack the sensitivity and specificity required for the detection of very early prostate lesions and diagnosis ultimately relies on an invasive biopsy. Once prostate cancer is diagnosed, there are no available prognostic markers for prostate cancer that provide information on how aggressively the tumor will grow. Therefore, more intrusive therapeutic routes are often chosen that result in a drastic reduction in the quality of life for the patient, even though the majority of prostate tumors are slow growing and non-aggressive. To identify prognostic biomarkers that can be used to molecularly distinguish patients with less aggressive tumors, we have collected data on DNA methylation patterns at more than 450,000 CpG loci in prostate tumor tissues and patient-matched normal prostate tissues using large-scale hybridization-based technology. We are currently performing reduced representation bisulfite sequencing (RRBS) on these prostate samples, which will provide DNA methylation status for an additional ~1,000,000 CpGs. Preliminary analysis has demonstrated that tumor and normal can be easily separated based on DNA methylation patterns and analysis is ongoing to identify CpGs that have prognostic value. We are presently sequencing RNA isolated from these prostate samples, which will provide detailed information on gene expression in both normal and tumor prostate tissue. We hope that the integration of these data sets with clinical follow-up data will allow us to identify candidate diagnostic and prognostic biomarkers on a genome-wide scale.